Analyzability and semantic associations in referring expressions : a study in comparative lexicology

Urban, M.

Chapter 5

Results I: Quantitative Evaluation

5.1. INTRODUCTION

This chapter is concerned with quantitative aspects of lexical motivation. It seeks to explain the behavior of the languages of the world with respect to their characteristic lexical profiles by asking questions such as: which languages have many morphologically complex terms and why? Are there languages that prefer metaphor-driven conceptualizations over contiguity-driven ones and why?

Since these are essentially quantitative questions, quantitative methods to analyze the data are needed, and therefore this chapter will make heavy use of statistics to come up with valid infererences and cross-linguistic generalizations. A concomitant and probably unavoidable effect is that for each language mostly abstract numbers rather than concrete lexical items, which ultimately are what can be and what is observed, will be analyzed statistically. In other words, there is a danger of tinkering statistically with numbers whose connection to the properties of actual languages is sometimes rather hard to see. This possible impression will be countered by making ample use of case studies that tie the data and the observed correlations to actual synchronic or diachronic observations about the languages in question to make the findings more palpable to the reader, and more generally to avoid the danger of an unduly abstract feel of quantitative analysis. Still, this chapter is characterized by quantitative methodology and probabilistic statements. The following chapter six, which is concerned with individual meanings and the cross-linguistic properties of the terms expressing them, will have a less quantitative and more of an anthropological orientation, and it will use ample data from individual languages.

5.2. DEGREE OF ANALYZABILITY: BASIC ANALYSES

In this section, the discussion of the different degrees of analyzable lexical items found in the languages of the world is entered. To give a first impression of the variability found here, the map in figure 1 shows the relative percentage of morphologically complex expressions of all languages in the core sample.

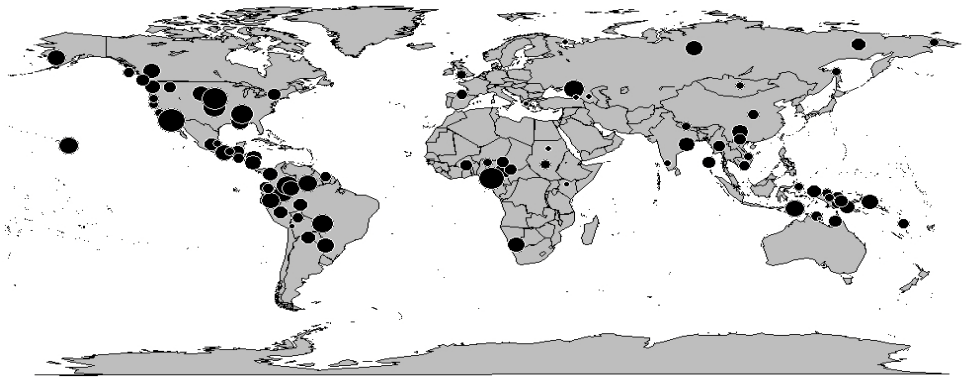


fig. 1: percentage of morphologically complex terms, core sample

The eye-catching areal clusterings will be discussed in § 5.3. But first, a basic comparison of the obtained values with that from another source, namely the World Loanword Database, follows as a kind of reliability check.

5.2.1. COMPARISON WITH THE DATA FROM THE WORLD LOANWORD DATABASE (HASPELMATH AND TADMOR 2009C)

The World Loanword Database (Haspelmath and Tadmor 2009c) contains vocabularies with about 1,000-2,000 entries for 41 languages of the world; the choice of meanings is based on, but not identical with Buck (1949). As the editors themselves note, there is a bias in the data towards European languages and thus the choice of languages is not necessarily representative of cross-linguistic diversity. The goal of the project is a systematic investigation of borrowability in different semantic domains and the varying degree of loanwords in different languages. Along with information on the status of each individual lexical item with respect to borrowing, contributors were asked to systematically code whether the lexical items are morphologically complex and, if they are, to provide a morphological analysis. This offers a convenient possibility for comparing the results of both investigations for each of the meanings that figure in both projects. This is measured by Haspelmath and Tadmor's "simplicity score," which is computed somewhat differently and thus requires some transformation to make the values comparable. The simplicity score, as the name suggests, measures morphological simplicity as opposed to morphological complexity, which is why it was converted into a measure of complexity for present purposes by subtracting the simplicity score from one. Further, the simplicity score of an individual lexical item is defined as being 1 for unanalyzable lexemes, .75 for semianalyzable lexemes and .5 for analyzable ones. To account for the difference in the scales and to convert the results into percentages the resuling value was multiplied with 200; in summary, the formula for converting simplicity scores is 200 × (1-simplicity score). Note, however, that the difference with respect to semianalyzable lexemes remains, since

they are assigned an intermediate value by Haspelmath and Tadmor while they are not taken into account in the present study at all.

129 of the 160 meanings presently investigated are also found in the World Loanword database, and the data show that the values of the different studies are often in close agreement to one another. Figure 2 is a scatterplot of the values obtained from both studies (data are in appendix D) that shows this correlation visually and is thus more accessible. Meanings with translational equivalents for only two or less languages in Haspelmath and Tadmor (2009c) were ignored in this plot; the values for 'lightning' and 'bolt of lightning' were averaged to a value of 43. As immediately becomes clear, there is a strong correlation: on average, the higher the value of morphological complexity for a given meaning in the present study, the higher the modified simplicity score value from Haspelmath and Tadmor (2009c). Unfortuntately, statistical testing is not permitted in this case, because the data overlap in some cases as some data from the World Loanword Database are included in the present sample. However, the close agreement between the two samples is very unlikely to be entirely caused by this data overlap, since the large majority of data in the sample for the present study do not come from the World Loanword Database.

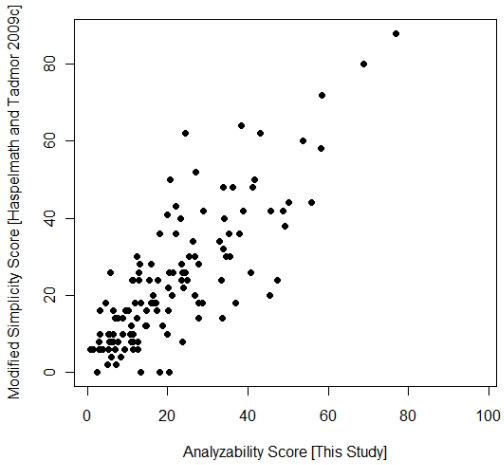


fig. 2: Correlation between modified simplicity scores from Haspelmath and Tadmor (2009c) and the measure of morphological complexity for each meaning that both studies investigate.

A further implication of this correlation pertains to the reliability of the data of this study: the vocabularies in the World Loanword Database are exclusively provided by experts on the respective languages, and thus it can be expected that all instances of morphological complexity were identified by the authors for the 41 languages in the database. As noted in Chapter 3 on methodology, mistakes in the recognition of morphological complexity in the present study cannot be ruled out and are indeed likely to occur to some degree, due to the necessary evil that the data are obtained mostly from secondary sources such as

dictionaries. However, the close agreement to the values derived from data provided by experts is a strong hint towards the assumption that overall, the data of the present study are by and large reliable.

Comparison with the World Loanword Database is also interesting in another respect: the wordlist of the World Loanword Database is considerably larger than the 160-item list of the present study. While the latter list, in spite of its relatively small size, offers a principled comparison of the sampled languages, it is nevertheless interesting to ask to what degree the values obtained from evaluating the data gathered for this list are representative for the situation with respect to the content-word inventory of the languages as a whole, i.e. including the verbal domain. Is the degree of analyzability here similar to that observed in the nominal domain? To tackle this question, Bradley Taylor has kindly computed the simplity score as defined in Haspelmath and Tadmor (2009c) for each of the languages in the World Loanword database by dividing the sum of simplicity scores for each word in a language by the sum of all words for that language. This value was transformed with the same formula as used above, and the resulting data are in Appendix B.

Notable is that the values from the World Loanword Database are significantly higher overall. This is likely a result of the fact that here many less "basic" meanings are taken into consideration which is a rough confirmation of the intuition that lexemes in more specialized vocabulary areas are more often morphologically complex than words for relatively basic concepts. However, while the values are consistently higher, the degree to which they are so vary significantly, from the rather modest difference of 5.22% in Takia up to more than 36% in Gurindji.[1] Figure 3 plots the results.

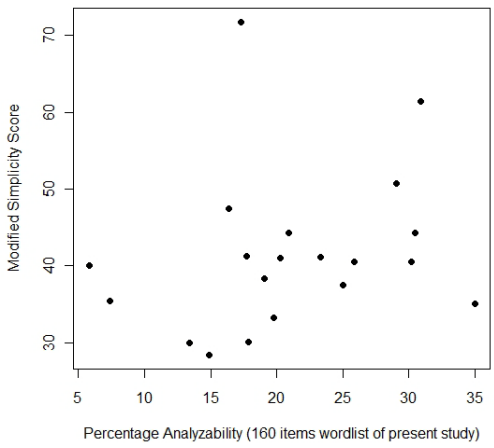


fig. 3: Correlation between modified simplicity scores from Haspelmath and Tadmor (2009c) and the measure of morphological complexity for each language that both studies investigate.

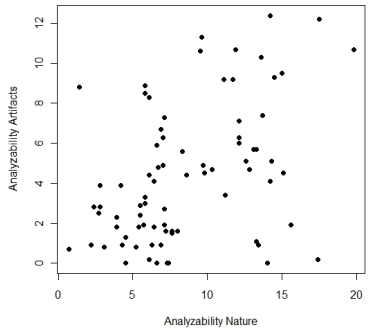
[1] Note that the even more extreme difference seen in the data for Mandarin are primarily due to the high amount of semantically redundant compounds, which are disregarded here, but which are counted in Haspelmath and Tadmor (2009c).

Again, since in this case the data for the 160-item list are a subset of the much larger overall vocabulary, statistical testing is not permitted. Figure 3 reveals a slight upward trend in the overall simplicity score as morphological complexity in the 160-items list increases, but there appears to be no strong dependency between the variables. In a way, this result is unsurprising, given cross-linguistic differences in the complexity of nouns and verbs reported e.g. for Kalam by Pawley (1993) and the typology of verbally and nominally oriented languages outlined by Talmy (2000: 59endnote11).

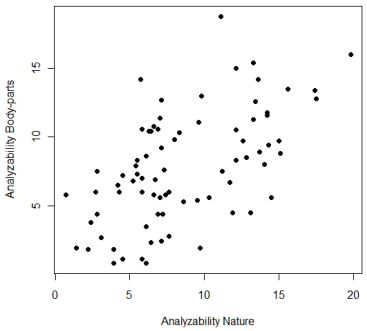
5.2.2. ARE THERE SIGNIFICANT DIFFERENCES IN ANALYZABILITY BETWEEN LANGUAGES ACROSS SEMANTIC DOMAINS?

It is conceivable that a language may rely on morphologically complex expressions in one semantic domain, while having an essentially unanalyzable lexicon in another. Data for percentages of analyzability for each language in the sample assessed over all semantic domains is in Appendix B, where information as to how the global value is distributed over the individual semantic domains is also provided (slight deviations from the global value are due to rounding).

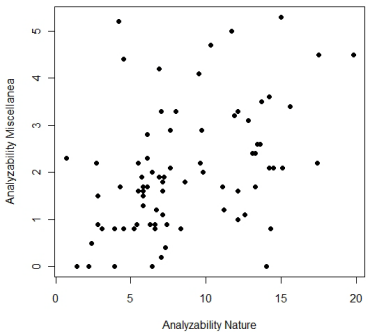
The question how morphological complexity is distributed across domains can be statistically assessed by performing correlation tests for each semantic domain with the others on the basis of the statistics sample. The diagrams in figure 4 plot the correlation between analyzablity in the four semantic domains; a correlation measure (Spearman's ρ) and an approximate *p*-value (due to ties) is provided in addition. The reported *p*-values are adjusted using the Bonferroni correction as implemented in R because of multiple testing.



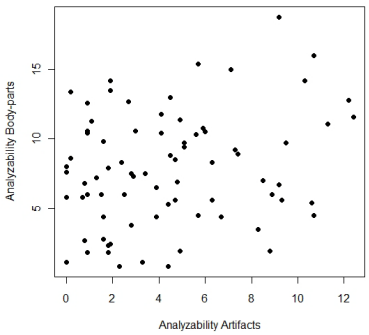
Nature vs. Artifacts: ρ ≈ .40, $p < .002$



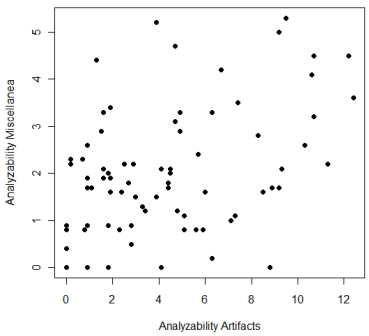
Nature vs. Bodyparts: ρ ≈ .54, $p < .0001$



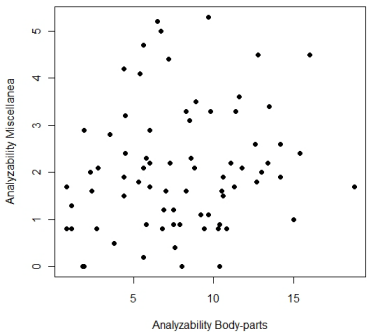
Nature vs. Miscellanea: $\rho \approx .44$, $p < .0004$



Artifacts vs. Bodyparts: $\rho \approx .20$, $p \approx .47$



Artifacts vs. Miscellanea: $\rho \approx .36$, $p < .007$



Bodyparts vs. Miscellanea: $\rho \approx .23$, $p \approx .28$

fig. 4: correlations in analyzability between meanings in different semantic domains

While there clearly is variation between semantic domains (this pertains in particular to the domain of artifacts and bodyparts), there are no dramatic cross-linguistic differences. The generalization that emerges is that THE ANALYZABLE TERMS ARE ON AVERAGE DISTRIBUTED FAIRLY EVENLY OVER THE SEMANTIC DOMAINS. The relative degree of analyzability can therefore be seen as a feature of the nominal lexicon as a whole, with differences found dominantly in the treatment of artifacts as items of acculturation on the one hand, and parts of the body on the other. There are languages where bodyparts are predominantly designated by morphologically simple lexemes on the one hand, but where the domain of artifacts, in contrast, is characterized by a high degree of analyzability, and vice versa. As will become clear in the following section, it is no surprise that artifact and body-part terms are the two semantic domains where no correlation in the overall analyzability is found, because there are areal differences in the distribution of analyzability in these domains.

5.3. ARE THERE AREAL FACTORS CONDITIONING THE DISTRIBUTION OF ANALYZABILITY?

It is conceivable that there are areal factors in play that govern the distribution of the prevalence of morphological complexity in the world's languages. When eyeballing the map plotting the world-wide distribution of the degree of morphologically complex terms in figure 1, some areal differences are apparent. For instance, there is a more or less contiguous area of low morphological complexity linking Eurasia (which also has some notable outliers such as Abzakh Adyghe, Ket, and Sora, discussion of which is in § 5.4.2.12.5.) with the North-Eastern part of Africa. Genealogically, it is interestingly the Afro-Asiatic languages in the sample, members of a family which is distributed over both Africa and Southwestern Eurasia through the Semitic branch, that pattern with Eurasia. Southeast Asia and Oceania appear to be areas of a moderate degree of morphological complexity overall, though there is some variation in particular in the New Guinea area. Marked differences again emerge in the Americas. In general, within the Americas, there is an West-East cline with respect to the variable, with lower values found in the Eastern part of North America. Likewise, in South America, languages of the greater Amazon region tend to score higher than languages spoken further in the West, in particular those spoken in an Andean environment. However, to really assess areality, mere eyeballing of maps is a dubious procedure (Cysouw 2005, Bickel and Nichols 2009), and statistical analysis is needed.

5.3.1. MORPHOLOGICAL COMPLEXITY, ALL DOMAINS

First, examining the percentages of analyzable terms in the whole set of 160 meanings under investigation, without recognition of differences that may exist with respect to the semantic domains investigated. there is no clear effect of area on the degree of analyzability using Dryer-6 (χ^2 = 10.1461, df = 5, p = .0712 by a Kruskal-Wallis rank sum test; all p-values in the further discussion of this section were obtained using this test). Under this breakdown, North American languages score very high. The corresponding plot is seen in figure 5.

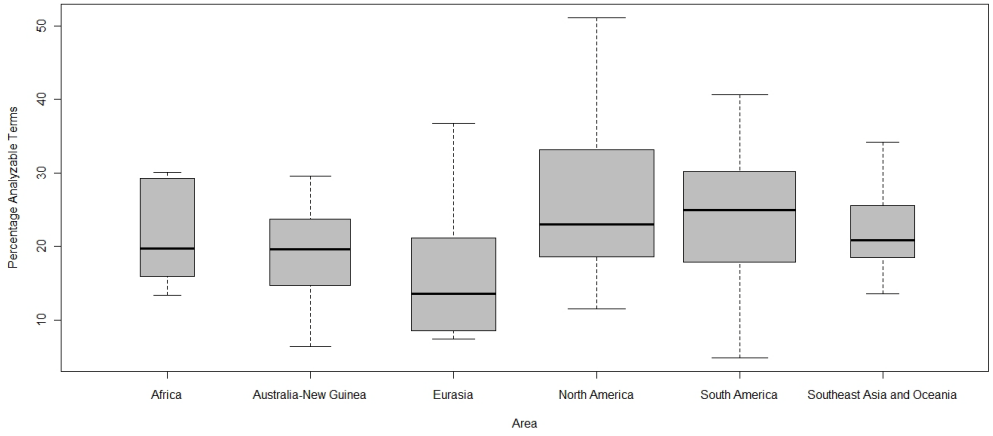


fig. 5: Areal breakdown of the degree of overall analyzability, using Dryer's (1992) breakdown

With the most fine-grained Nichols-11 breakdown, the differences between areas is closer to significance (χ^2 = 17.8067, df = 10, p = .05831). What this plot shows (and what is also suggested by impressionistically eyeballing the map) is that languages of Eastern North America score very high. Figure 6 plots the results.

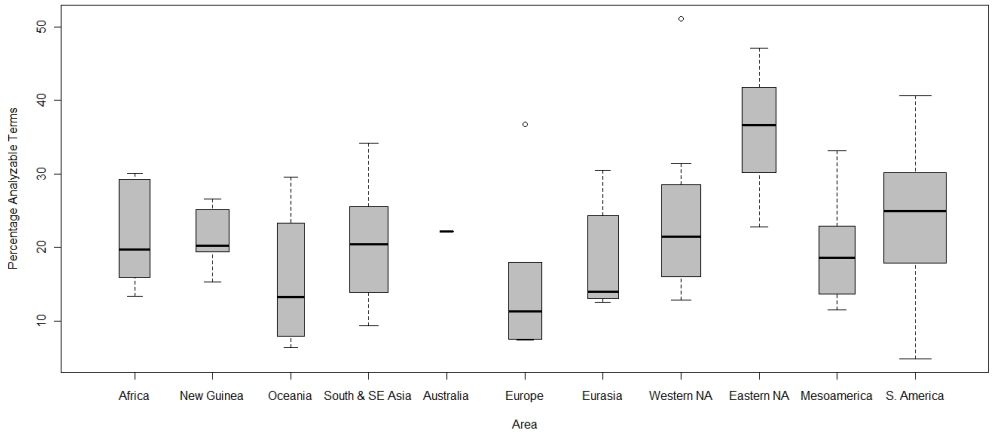


fig. 6: Areal breakdown of the degree of overall analyzability, using Nichols's (1992: 25-26) breakdown

An even stronger statistically significant difference emerges when using the broadest of the three partitionings: moving from the Old World via the Pacific into the New World, the

degree of overall analyzability rises significantly (χ^2 = 9.6076, df = 2, p = .008199), as visualized in figure 7.

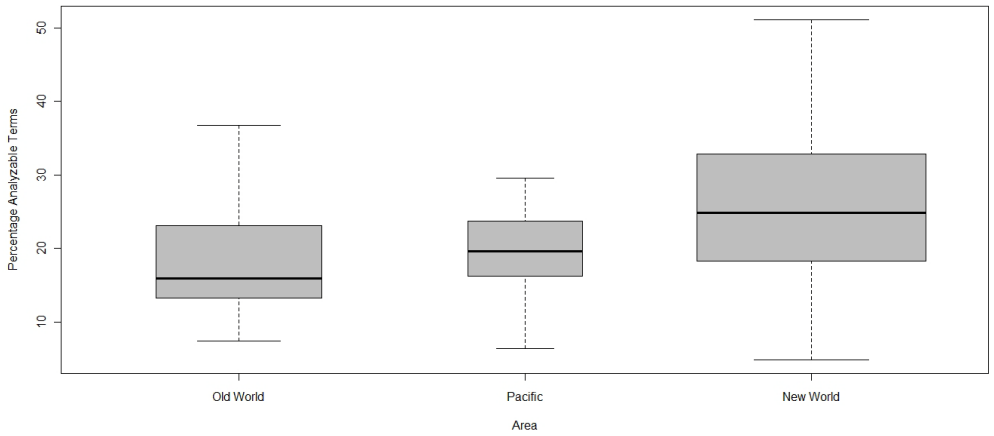


fig. 7: Areal breakdown of the degree of overall analyzability, using Nichols's (1992: 27) breakdown.

While this is an interesting result, it also raises a question, namely whether the difference is due to historical contingency rather than to language-inherent properties, as most of the modern-day artifacts were invented in the Old World and are relatively recent newcomers in many parts of the World. Consequently, it would make sense to expect that Old World-terms for artifacts are often unanalyzable due to their age, whereas neologisms in the New World have a clearly discernible morphological structure, as the artifacts have only be known for a short time span. Therefore, the same tests were performed, but with removing data from the artifact category from the data pool. If there remains a correlation, this would be an indication of genuine macro-areality. Under these testing conditions, the tendency for areality when using the Dryer-6 and Nichols-11 breakdowns ceases (χ^2 = 9.0625, df = 5, p = .1066 and χ^2 = 12.4319, df = 10, p = .2572 respectively). Although the same basic difference between the Americas on the one hand and other areas of the world, in particular Eurasia remains, this difference is not significant.

The statistical correlation with the three broadest possible sample areas as used in the Nichols-3 breakdown is weakened to p = .0336 (χ^2 = 6.7867, df = 2), with the ranking in the degree of analyzability from the Old World via the Pacific to the New World remains intact, as seen in figure 8. Thus, when artifact terms are not taken into consideration, there is NO CLEAR AREAL EFFECT ON THE DISTRIBUTION OF OVERALL MORPHOLOGICAL COMPLEXITY IN THE INVESTIGATED SLICE OF THE LEXICON UNDER THE SPLIT-UPS USED FOR TESTING. This should not necessarily be taken to entail that there cannot be areality on a smaller scale (cf. Bright 2004, tentatively also Nichols and Nichols 2007); to assess these, however, a much larger sample size would be needed. The findings should also not be interpreted in the sense that the semantic structure of analyzable terms as well as in colexification is not sensitive to lan-

guage contact and thus to areal effects (see extensive discussion in § 6.4.3). When it comes to sheer quantity of analyzable terms, however, areal factors appear to play an at best subordinate role.

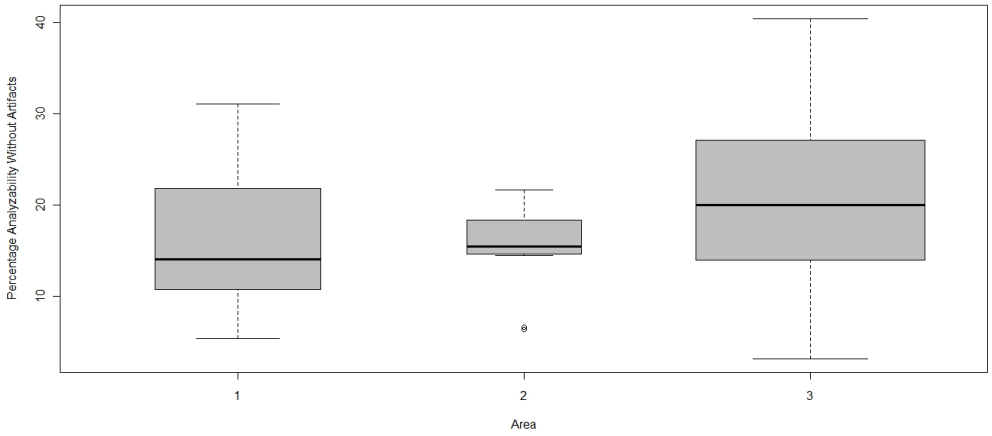


fig. 8: Areal breakdown of the degree of overall analyzability with artifacts removed, using Nichols's (1992: 27) breakdown.

Given that even large linguistic areas, if they exist, are the outcome of language contact, this is an indication that there seems to be little pressure on languages in contact to adjust the morphological structure of their lexicon (preponderance for complex terms in general on the one hand or preponderance of simplex lexical items on the other) to each other. Of course, to reiterate, this does not entail that calquing of morphologically complex expressions for a given referent does not occur – it does, but on a large scale, at the level of the lexicon at large, such tendencies seem to be rather weak.

5.3.2. Individual Semantic Domains

This section assesses differences in analyzability in the four semantic domains used in this study, with the same three breakdowns used for testing. There is no appreciable difference in analyzability of nature-related and topological terms under all three breakdowns (Dryer-6: $\chi^2 = 7.3432$, df = 5, $p = .1963$, Nichols-11: $\chi^2 = 7.1813$, df = 10, $p = .7082$, Nichols-3: $\chi^2 = 2.634$, df = 2, $p = .2679$). In contrast, as one would expect from the exercise of removing the artifact domain from the global calculations above, there is an areal effect on the analyzability of artifact terms under the Nichols-11 and Nichols-3 breakdown ($\chi^2 = 19.108$, df = 10, $p = .03891$ and $\chi^2 = 9.3812$, df = 2, $p = .009181$ respectively). The associated plots are seen in figures 9 and 10.

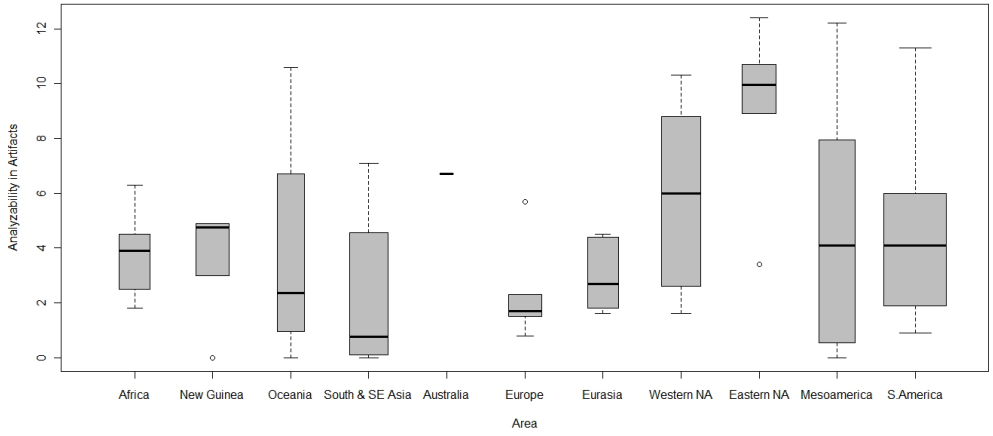


fig. 9: Areal breakdown of the degree of overall analyzability in artifacts, using Nichols's (1992: 25-26) breakdown.

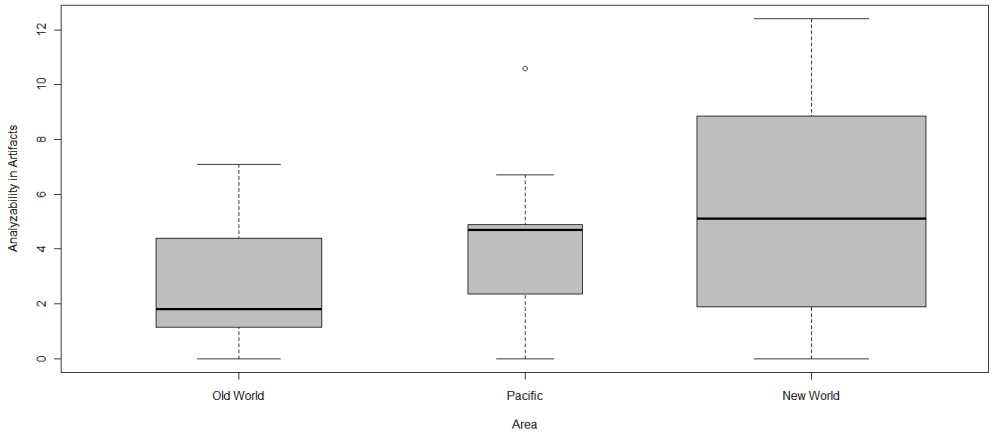


fig. 10: Areal breakdown of the degree of overall analyzability in artifacts, using Nichols's (1992: 27) breakdown.

The same tendencies – higher degrees of analyzability in artifacts in the Americas and very high degree of analyzability in artifact terms in North America – discerned by the application of the aforementioned breakdowns emerges when testing for Dryer-6, although the result is not quite significant (χ^2 = 10.0677, df = 5, p = .07334). A simple and straightforward conclusion follows, although it is hardly suprising: ANALYZABLE TERMS FOR ARTIFACTS ARE FOUND AT A HIGHER RATE IN THOSE AREAS OF THE WORLD WHERE THEY ARE RECENT ITEMS OF ACCULTURATION, and this notwithstanding the fact that another obvious option for lexical acculturation is borrowing of a word for a novel artifact from a contact language (this is

further discussed in § 5.4.2.7.1.). This obviously is an instance of what Haiman (1985: 149) calls the "iron horse" effect: "Languages tend to have complex periphrastic means of expressing notions that are unfamilar."

Moving on to the domain of body parts and body fluids, again there are differences between the areas tested under the Dryer-6 breakdown (χ^2 = 12.5949 df = 5, p = .02749). As seen in the corresponding plot in figure 11, it is South American languages that on average have the highest degree of analyzability in this domain.

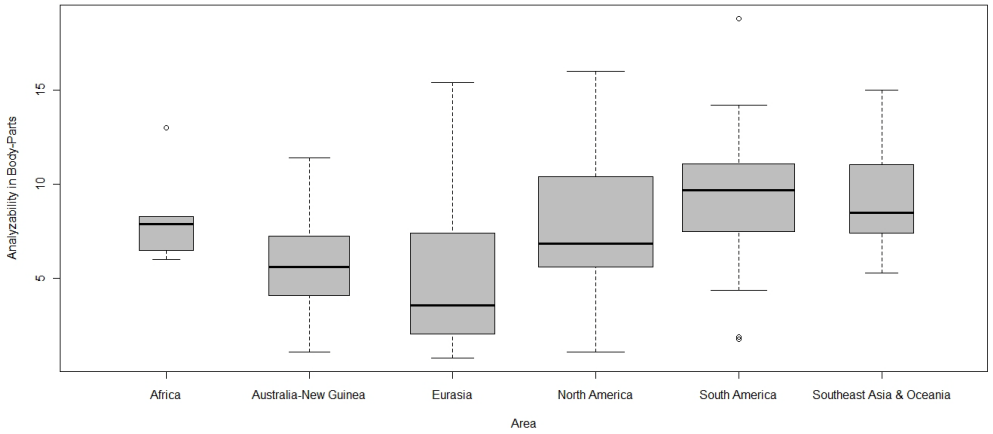


fig. 11: Areal breakdown of the degree of overall analyzability in body-part and body-fluid terms, using Dryer's (1992) breakdown.

The sharpest contrast is that between Eurasia, where body-part terms are least frequently analyzable and the Americas, in particular South America, where they are on average most commonly so. This is mirrored by the results of applying the Nichols-3 breakdown: as for the global values, there is a cline of rising degrees of analyzability moving from the Old World via the Pacific into the New World. The test for the Nichols-11 and Nichols-3 breakdowns are, however, not significant statistically (χ^2 = 15.0744, df = 10, p = .1294 and χ^2 = 4.9069, df = 2, p = .086 respectively), but also in the former, the high degree of analyzable body-part terms in South America is noticeable.

The difference between South America and the rest of the world in the evaluation based on Dryer-6 is mild, and may be due to two unrelated factors, namely the common process of derivation of body-part terms via sortal classifiers in a number of languages of the broader Amazon region (see § 4.4.1. for details), as well as a general increased presence of analyzable body-part and body-fluid terms, most of which are not particularly uncommon in the rest of the world in their semantic structure. Examples of body-part and body-fluid terms involving a sortal classifier from Bora (see Seifart 2005 for discussion of these in the Miraña dialect) are in (1.).

(1.) a. *nijpá-yu* urine-CL.ROUND 'bladder'
b. *máátyo-u* crying-CL.ROUND 'tear'

As for other complex terms in the domain of body-part terms, some languages of South America are unusual in that they have analyzable terms for 'mouth' (e.g. Tsafiki *fi'quí foró* 'language opening/hole') and 'stomach' (see Appendix E, 124 and 138). Also common are complex terms for 'vein,' most often via a metaphor involving either 'way, road,' such as Huambisa *numpa jinti* 'blood way' or sometimes on the basis of 'liana' (the conceptualization via 'way, road' is also heard of in other regions of the world).

There is no discernible areal effect when testing the domain of phases of the day and miscellanea (Dryer-6: $\chi^2 = 4.1783$, df = 5, $p = .524$, Nichols-11: $\chi^2 = 10.3555$, df = 10, $p = .4099$, Nichols-3: $\chi^2 = .8169$, df = 2, $p = .6647$) – an unsurprising result, given the heterogenous nature of this group of vocabulary items. Worth noting in this context is also the absence of areal effects on the overall degree of nature-related and topological terms, because also this group of meanings is fairly heterogenous. While the meanings clearly can be subsumed under a common denominator, it is still the case that they may be broken down into several smaller subdomains, such as the conceptualization of bodies of water, of things that have to do with fire, the heavenly bodies, parts and products of animals, etc. However, it is far from clear whether they form a lexical field that has the same degree of conceptual coherence that the domains of artifacts and body-parts possess. The results thus open up the possibility that it might be well-circumscribed semantic domains such as the body-part vocabulary and artifacts (demonstrated in Cognitive Psychology by priming experiments e.g. by Neely 1977 and Moss et al. 1995, in the case of artifacts assisted by historical contingencies), rather than the lexicon in general, which are likely to host areal clusterings of morphologically complex terms.

5.3.3. SUMMARY

Summing up, in the assessment of possible areality in the overall degree of morphological complexity, a statistical trend for languages in certain areas can be noted that, however, is so mild that one cannot discern a clear areal effect. Closer inspection of the individual semantic domains under investigation revealed that the degree of analyzability in artifacts and to a lesser degree in body-part and body-fluid terms is unequal in different areas of the world. Importantly, these differences mirror the general trend when evaluating overall vocabulary – rising degree of analyzability when moving from the Old World to the New World, in particular (parts of) North America. In effect, it appears that the structure of the vocabulary for body-parts and artifacts is at large responsible for the trend that is observable on a global scale, while the domains of nature-related and topological terms and phases of the day and miscellanea weaken it.

5.4. ANALYZABILITY IN THE LEXICON: TYPOLOGICAL PERSPECTIVES

5.4.1. STRUCTURAL CORRELATIONS TO THE DERIVED-LEXICAL CONTINUUM?

In § 4.7., a correlation between a preponderance of derived terms and the elaborateness of verbal person marking was established. Taking up this thread, this section explores whether there are further structural features that correlate with this distinction as to the type of analyzable lexical items. Since it is hitherto at large unclear what, if any, further factors may be relevant here, correlation tests using the data for the features in the World Atlas of Language Structures (Haspelmath et al. 2005) were performed. These tests are meant to be hypothesis-generating rather than hypothesis-testing. For that reason tests were carried out for the entire set of WALS features, regardless of how unlikely a connection between a given feature and the distinction between complex lexical items of the derived and lexical type may seem.

A word of caution in the interpretation of the findings is in order. While all languages in the statistics sample of the present study are also featured in WALS, it is not necessarily the case that very many datapoints are coded for them. While for Basque, for instance, a value is coded for 127 out of 138 features, a value for a meager eight features is available for Berik. In other words, it is the case for many features that the datapoints available for statistical testing are greatly reduced due to lack of coding in WALS (or the grammatical descriptions such coding presupposes), and in turn, the reliability of any statistical test depends to some extent on the available amount of coded data. Thus, the search for typological correlations on the basis of WALS need to be regarded as preliminary, in particular where the empirical database is small (see Wohlgemuth 2009: 187-189 for similar discussion). The preliminary tests on WALS yielded significant correlations with as many as ten WALS features:

(i) Imperative-Hortative systems (Van der Auwera et al. 2005):
$\chi^2 = 7.4559$, df = 3, $p = .0587$, Kruskal-Wallis Rank Sum Test
(ii) Order of Subject, Object, and Verb (Dryer 2005g):
$\chi^2 = 13.6505$, df = 5, $p = .01799$, Kruskal-Wallis Rank Sum Test
(iii) Order of Subject and Verb (Dryer 2005f):
$\chi^2 = 9.8122$, df = 2, $p = .007401$, Kruskal-Wallis Rank Sum Test
(iv) Order of Object and Verb (Dryer 2005d):
$\chi^2 = 5.598$, df = 2, $p = .06087$, Kruskal-Wallis Rank Sum Test
(v) Order of Adjective and Noun (Dryer 2005b):
$\chi^2 = 9.6764$, df = 2, $p = .007921$, Kruskal-Wallis Rank Sum Test
(vi) Position of Polar Question Particles (Dryer 2005i):
$\chi^2 = 8.3482$, df = 4, $p = .07963$, Kruskal-Wallis Rank Sum Test
(vii) Position of Interrogative Phrases in Content Questions (Dryer 2005h):
$\chi^2 = 8.2179$, df = 2, $p = .01642$, Kruskal Wallis Rank Sum Test
(viii) Relationship Between the Order of Object and Verb and the Order of Adjective and Noun (Dryer 2005j):
$\chi^2 = 11.4475$, df =4, $p = .02197$, Kruskal Wallis Rank Sum Test
(ix) Verbal Person Marking (see § 4.7)

(x) Nonperiphrastic Causative Constructions (Song 2005):
$\chi^2 = 5.5664$, df = 2, p =.06184, Kruskal Wallis Rank Sum Test

There are many features pertaining to word-order typology that yield significant or near-significant p-values, one is the overall classical Greenbergian word order typology, among the others are those looking at the order of subject and verb, the order of object and verb, and the order or adjective and noun specifically. However, as is well known, word order patterns are subject to areal pressure; in fact, the example of basic word order was the very trigger in linguistic typology to recognize that areal factors need to be taken into account when searching for universals in the classical sense (Dryer 1992). It is thus especially imperative to control for areal factors in the final analysis, using a Linear Mixed Effects Model (see § 4.7. for details), with the hypothesis to be tested in each case being that there indeed is a genuine influence of the above features on the degree of analyzability. Six of the above features (next to verbal person marking already discussed in chapter 4) survived closer scrutiny when controlling for areal effects[2] by employing the Mixed Model design familiar by now (in all models, the percentage of derived terms was square-transformed; for the feature concerning the order of subject and verb, no model could be built because even after various transformations residuals were still not normally distributed and the resulting model was therefore not valid):

(i) Order of Subject, Object and Verb (Dryer 2005g): p =.0247
(ii) Order of Object and Verb (Dryer 2005d): p =.0298
(iii) Order of Adjective and Noun (Dryer 2005b): p =.0053
(iv) Position of Interrogative Phrases in Content Questions (Dryer 2005h): p =.0018
(v) Relationship between the Order of Object and Verb and the Order of Adjective and Noun (Dryer 2005j): p =.0174
(vi) Nonperiphrastic Causative Constructions (Song 2005): p =.04003

The boxplot in figure 12 shows the distribution of the sampled languages with regard to the percentage of derived terms depending on the possible orders of subject, object and verb in the original statistics sample.

[2] Note that in principle the reverse situation, namely that effects only become visible rather than disappear when taking into account areal factors, is also conceivable. Since models have not been constructed for each of the WALS features, it is possible that there are some undetected WALS features for which a genuine correlation might exist.

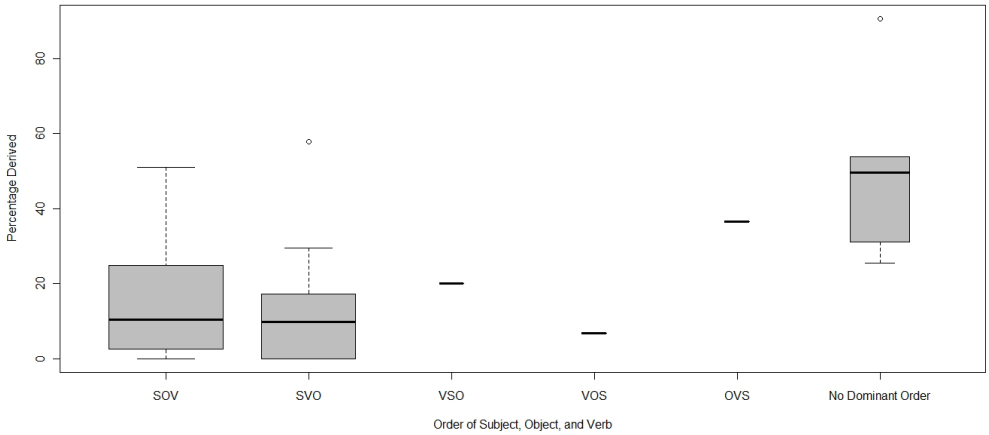


fig. 12: Percentage of derived terms depending on word order typology

As figure 12 shows, the main difference is not between languages with a fixed preferred word order of some kind, but rather between these and languages in which no particular grammatically conditioned word order is dominant.

The same basic observation can be made for the order of object and verb: here, too, it is the languages without dominant order that stand out in featuring an elevated number of derived terms, as seen in figure 13.

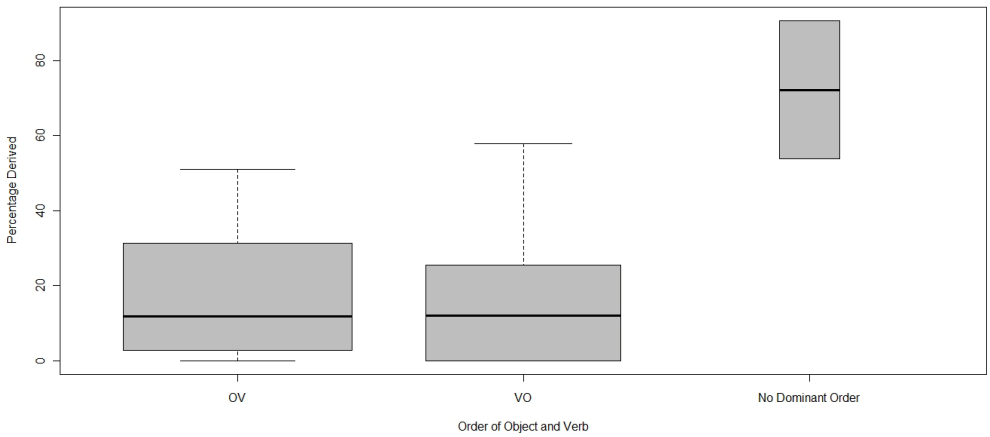


fig. 13: Percentage of derived terms depending on the order of object and verb

With respect to the order of adjective and noun, again it is the language where the order of these elements is not fixed that score highest with regard to the percentage of derived terms (figure 14).

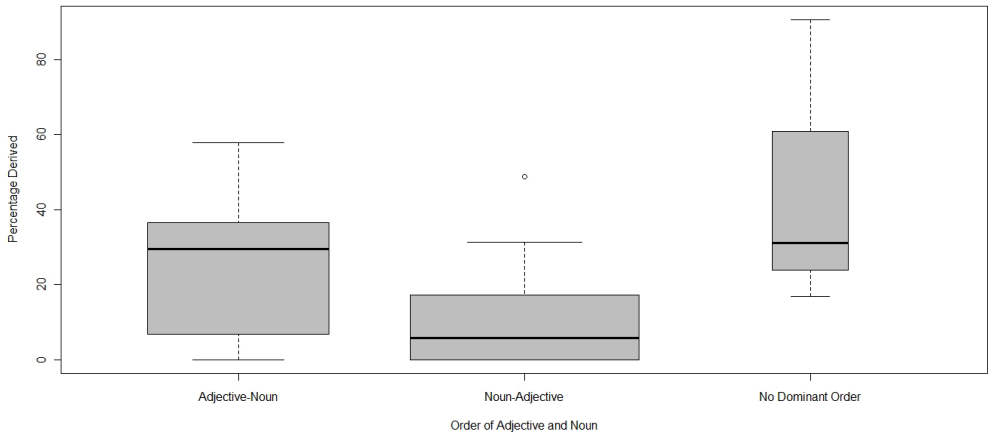


fig. 14: Percentage of derived terms depending on the order of adjective and noun

Bearing in mind the significant correlation with verbal person marking established in § 4.7, the trends seen so far seem easily accountable for: if information as to the arguments is coded morphosyntactically on the verb, there is functionally little need for fixed word order to make clear who does what to whom. Further, the correlation would also be additional evidence for a particular typological profile favoring terms of the derived type with synthetic morphology and concomitantly free word order.

Moving on to other significant correlations, as the boxplot in figure 15 shows, there is a drop in the percentage of derived terms in languages with non-initial interrogative phrases as opposed to those with initial interrogative phrases.

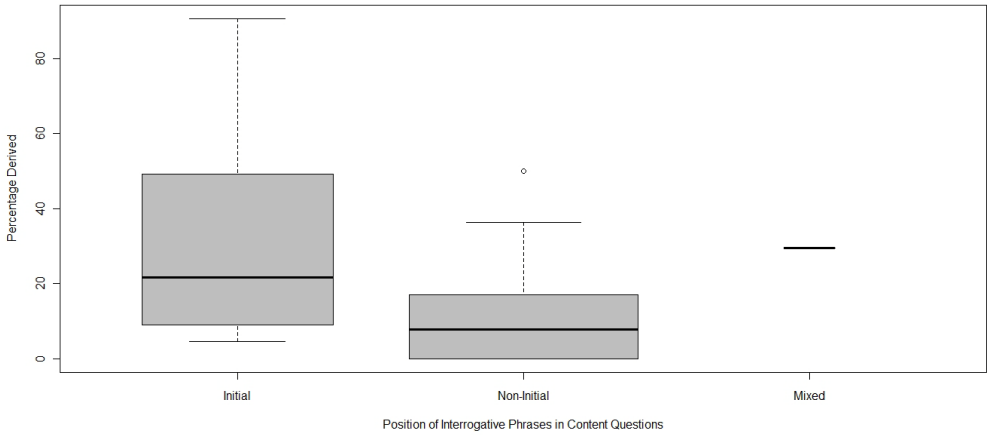


fig. 15: Percentage of derived terms depending on the position of interrogative phrases in content questions

Another significant correlation that is also independent of the basic word order typology according to Dryer (2005j) is that concerned with the order of object and verb on the one hand and that of the order of adjective and noun on the other. Notably, both variables on their own yielded significant interactions, as has already been discussed. Consistent with the findings made there, it is also here the languages grouped in the category showing an "other" behavior than the four logically possible main types (figure 16), that is, in Dryer's (2005j) coding, such languages where either or both order of object and verb or adjective and noun is not fixed or where constructions modifying nouns with adjectives are absent.

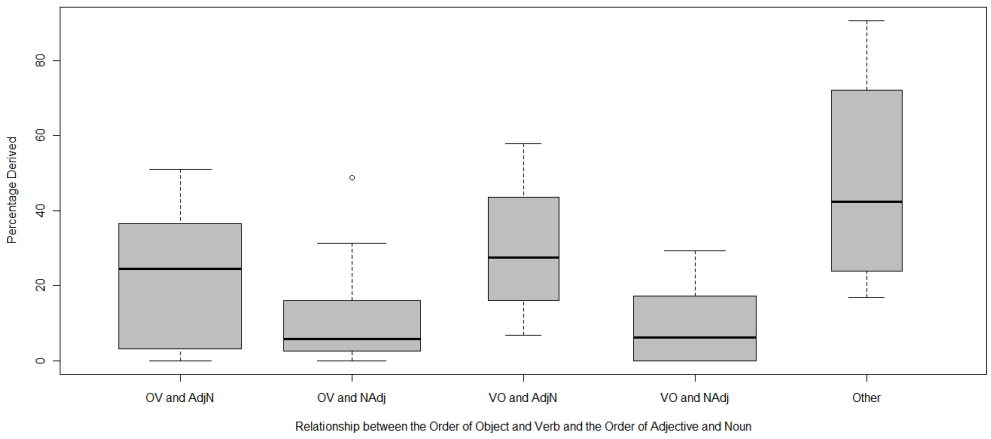


fig. 16: Percentage of derived terms depending on the order of object and verb and the order of adjective and noun

However, there still is one issue: ultimately, the initial tests leading to each of the hypotheses were part of a very large series of exploratory tests on the entire WALS dataset (not corrected for multiple comparison, as suggested for exploratory investigations by Bender and Lange 2001). Given the fact that for each test, there is a chance of $\alpha = .05$ percent that a significant result is obtained in the absence of any real effect, one can expect a number of about 7 tests with spurious significance simply due to chance. Therefore, it is furthermore imperative to cross-validate the results. Fortunately, this is possible, since data for more languages than those in the statistics sample were collected. From the remaining languages for which data is available for 65 percent or more of the investigated meanings, a genealogically balanced VALIDATION SAMPLE, as alluded to in § 3.3., was constructed. This includes the following languages, chosen randomly if more than one option was available for a particular language family: Swahili, Kanuri, Dongolese Nubian, Burarra, Kosarek Yale, Greek, Japanese, Vietnamese, Blackfoot, Comanche, Kashaya, Wintu, Yuki, Tuscarora, San Lucas Quiaviní Zapotec, Huambisa, Wayampi, Tsafiki, Ancash Quechua, Kapingamarangi, Mandarin, and Lesser Antillean Creole French. If a correlation is genuine, one should be able to replicate the results on data from entirely different languages, and hence also on those in the validation sample. Mixed Models were constructed for all of the features in the above list of six features. The estimates for the fixed effects were compared with those from the original models, and the correlation was taken to be genuine if they are within the range of the original estimate ± its standard error.

As for the exceptional behaviour of languages without fixed word order in the overall typology of order of subject, object, and verb, the estimate for the original sample is 3.7132 ± .9669, while that for the validation sample is only .2735, thus showing the same positive direction, but much more mildly and not within the range defined by the standard error around the estimate of the original sample. Hence, the effect must be rejected. The same is true for the subtypology looking only at the order of object and verb: the estimate from the original sample is 4.5910 and the standard error 1.7430, while the estimate for the validation sample is only .02062. For the correlation with the order of adjective and noun, validation is not possible because there is no language without a dominant order in the validation sample (the drop in the percentage of derived terms in languages with noun-adjective order present in the original sample, at any rate, cannot be replicated: original estimate is -1.9677 ± .7093 as opposed to -.2866 in the validation sample).

The correlation that can be most clearly replicated is the one which is at the same time most difficult to give reasons for, namely the position of interrogative phrases in content questions. The estimate for the difference in derived terms between languages with non-initial interrogative phrases and those with initial interrogative phrases from the original sample is -2.4418 ± .7494, while that of the validation sample is -1.8751, thus within the limits defined by the standard error (again, no evaluation of the behavior of languages with mixed position is possible since this group is very small and there are no representatives of it in the validation sample). Why this is the case is unclear; Dryer (2005h) does not mention correlations of this variable with other properties pertaining to word order, so that this feature seems unlikely to be a side-effect of a more easily explainable property.

With regard to the feature looking at the order of object and verb in relation to the order of adjective and noun, most estimates can be roughly replicated but notably not the most interesting one, namely the rise in derived terms in languages with a relationship other than the four major typological groupings recognized (estimates: -1.2492 ± .8800 vs. -0.39828, 1.1473 ± 1.3118 vs. -.08885, -1.8195 ± .9276 vs. -.42206, but 2.6139 ± 1.3118 vs. .57090).

The last of the significant correlations listed above, that with periphrastic causative constructions, is clearly disconfirmed by the evidence of the validation sample, at least for the group represented in both samples, namely morphological but no compound constructions (-1.298 vs. 4.86 ± 1.741).

Taken together, the results are suggestive, but the evidence from the validation sample suggests that the effect of word order typology, in particular the effect of free as opposed to fixed word order, is overestimated in the original sample and cannot at present be accepted as valid, while less obvious parameters of word order appear to have a replicable effect. Thus, verbal person marking seems to be the clearest correlate to the derived-lexical continuum that can safely be identified and at the same time explained functionally at present (which, of course, does not entail that it is the only one). Although the results are relatively meagre, the section at least serves to introduce the step-wise procedure used here to arrive at reliable correlations, and it will be made use of again in the following section, which approaches the question as to structural correlations to the degree of analyzability itself.

5.4.2. OVERALL MORPHOLOGICAL COMPLEXITY

5.4.2.1. Preliminary tests on the basis of WALS

This section seeks to elaborate on possible correlations between the degree of morphological complexity in the nominal lexicon as a whole and other typological properties of the sampled language, thus forming the major part of the entire chapter. The method employed is the same here as above: preliminary hypothesis-generating tests on the basis of WALS, elaborated on by more fine-grained analyses. Below are significant or near-significant correlations obtained by the preliminary tests.

(i) Consonant Inventories (Maddieson 2005a):
S = 10813.58, p = .01815, Spearman's ρ = -.391709
(ii) Consonant-Vowel Ratio (Maddieson 2005b):
χ^2 = 9.4684, df =4, p = .0504, Kruskal-Wallis Rank Sum Test
(iii) Syllable Structure (Maddieson 2005d)
S = 7627.729, p = .02406, Spearman's ρ = -.3980442
(iv) Possessive Classification (Nichols and Bickel 2005c):
S= 1026.556, p < .0001, Spearman's ρ = .6866435
(v) Semantic Distinctions of Evidentiality (de Haan 2005):
χ^2 = 9.8448, df = 2, p = .007282, Kruskal-Wallis Rank Sum Test
(vi) Order of Adjective and Noun (Dryer 2005b):
χ^2 = 6.5014, df = 2, p = .03875, Kruskal-Wallis Rank Sum Test

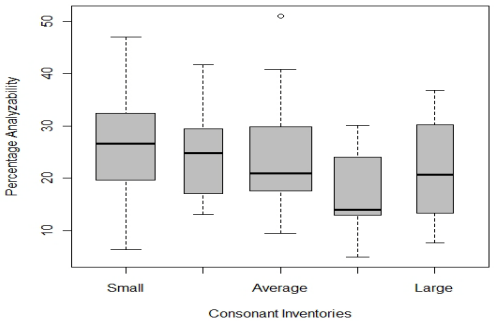
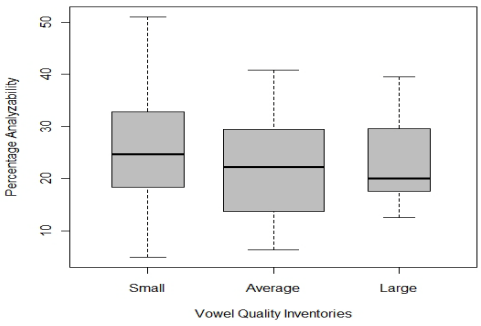
(vii) Order of Demonstrative and Noun (Dryer 2005c):
$\chi^2 = 8.8377$, df = 4, $p = .06529$, Kruskal-Wallis Rank Sum Test
(viii) Predicative Adjectives (Stassen 2005a)
$\chi^2 = 5.9285$, df = 2, $p = .0516$, Kruskal-Wallis Rank Sum Test
(ix) Purpose Clauses (Cristofaro 2005)
$\chi^2 = 3.0855$, df = 1, $p = .079$, Kruskal-Wallis Rank Sum Test

5.4.2.2. Elaborating on the preliminary findings with regard to phonology

Surprisingly, two phonological features are tested positively for significant interaction on the basis of the WALS data, and another one yields borderline significance. Apparently, the smaller the consonant inventory and the simpler the structure of the maximal syllable, the higher the amount of morphologically complex lexical items will be. Given that two features in the area of phonology yield significance, some real interaction is likely to go on between phonology and lexicon.

However, as noted above, results need to be interpreted with caution at this stage since there are many gaps in the data. In order to arrive at reliable results, and to examine whether the correlation can be substantiated, the policy adopted here is to amend the WALS database with data from published materials for the relevant phonological features in order to fill gaps in cases where the statistical testing on the basis of the WALS data revealed significance and the correlation appeared to be amenable to meaningful interpretation. In doing so, additional data were also gathered for the other pertinent feature in this area, namely vowel quality inventories, although here only a very weak negative correlation (Spearman's $\rho \approx -.08$) that is clearly not significant statistically ($p = .6368$) was found when testing on the limited WALS data. Information from published materials was coded in precisely the same fashion as in the relevant WALS features (Maddieson 2005a, b, d, h) to ensure compatibility of the data; furthermore, phonemes indicated to be non-native and restricted to loanwords were not counted in making coding decisions. Data are in appendix C. A problem for analysis is that the phonological features are highly unevenly distributed areally as revealed by Kruskal-Wallis rank sum tests (and as also suggested by Maddieson 2005a, d, h).

It is thus again particularly important to control for areal factors in the final analysis, using Linear Mixed Effects Models. The findings on the basis of the enhanced datasets for phonology are seen in table 1.

Feature	*p*-value	Plot
1. Consonant Inventories[3]	*p* =.0234 estimate: -1.977	
2. Vowel Quality Inventories	*p* = .5896, estimate : -.9965	

[3] An apparent clerical error in Maddieson (2005a) was corrected before performing analysis: Oneida, according to one of the sources consulted by Maddieson (Abbott 2000), should be, with nine distinctive consonant phonemes (/l/, /w/, /y/, /n/, /t/, /k/, /s/, /ʔ/, and /h/), coded as having a small, not moderately large consonant inventory. This coding decision would be valid even if one recognized voicing as distinctive in the alveolar stop as proposed in some analyses, but not followed by Abbott (2000).

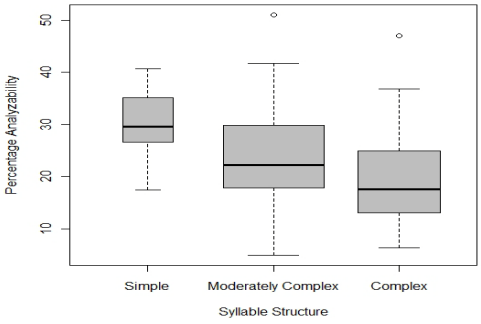
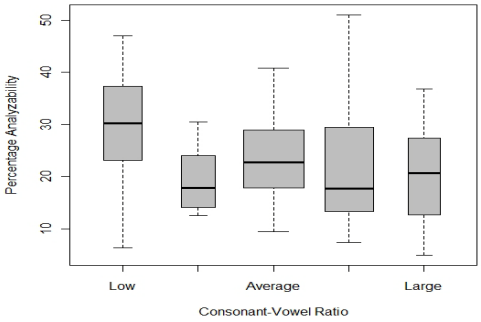
3. Syllable Structure[4]	p =.0102, estimates: -7.688, -13.053)	
4. Consonant-Vowel Ratio	p =.0401, estimate: -2.028	

table 1: Differences in the degree of analyzability depending on phonological factors, computed on an extended dataset based on WALS

Testing on the amended datasets substantiates the interaction between consonant inventories (the correlation is now a little weaker here, but the distribution is much more even and observations in each group sufficiently large) and syllable structure, and also confirms the insignificance of vowel inventory size on the degree of analyzability.[5] A correlation that is a bit difficult to interpret is that regarding the consonant-vowel ratio. This is calculated by simply dividing the number of distinctive consonants by the number of distinc-

[4] Maddieson's (2005d) coding decision was revised with respect to Tetun, which is coded by him as having a complex syllable structure, presumably because of Morris (1984) mentioning weakly articulated excrescent consonants in the syllable onset in emphatic speech. In spite of this, Tetun was coded as having a moderately complex syllable structure given Van Engelenhoven and Williams-van Klinken's (2005) description of Tetun syllable structure as (C)V(C).

[5] Note that also for this feature, it is true that languages with smaller inventories tend to have more analyzable terms. However, unlike for the other features, there are areal factors in play: when not controlling for area, a borderline significance emerges also for this feature. When areal differences are taken into account in the Mixed Model, significance ceases, so that it is not a valid cross-linguistic generalization to say that there is a direct influence of vowel inventory size on analyzability in the lexicon. This example underscores the importance of taking into account areal biases when formulating cross-linguistic generalizations.

tive vowel qualities. Ultimately, this entails that a languages with both few consonant and vowel phonemes and languages with both very many consonant and vowel phonemes will end up receiving similar scores in the ratio of consonant to vowels, and thus, this measure is in principle no measure of phonological complexity per se. However, it is important that the variance within consonant inventories is much greater than that within the vowel inventory system: while the number of distinctive consonants in Maddieson's (2005a) sample ranges from six to 122, the number of distinctive vowel qualities varies only between two and fourteen (Maddieson 2005h).[6] An effect of this is that, as noted by Maddieson (2005b), languages with large consonant inventories typically also have a large consonant-vowel ratio. Thus the areal distribution of the figure for consonant-vowel ratio sometimes overlaps with that for consonant inventories. This is noticeable for instance in the American Northwest. Many languages spoken in this region have both large consonant inventories and a high consonant-vowel ratio, whereas in Eastern South America, many languages have small consonant inventories and also a low consonant-vowel ratio. This at first glance somewhat hidden dependency is likely the key to explain why a significant correlation between the consonant-vowel ratio and the degree of analyzability is found.

Cross-validating the results using the validation sample already used above after amending data also for the languages in this sample (see appendix C for data), it turns out that the estimate for Consonant Inventories as a predictor in the validation model is -.1140, thus within the limits of that for the statistics model ± its standard error (-1.9768 ± .8416), and also well within these limits for the estimates for syllable structure (-4.292, compare -7.688 ± 3.486 and -12.952, compare -13.053 ± 3.990 respectively) and for the Vowel-Consonant-Ratio (-2.678, compare -2.0282 ± .9556). Hence, all correlations appear genuine.

To sum up, the smaller the consonant inventory of a language, the simpler the maximal syllable (and the lower the consonant-vowel ratio), in short, THE SIMPLER THE PHONOLOGICAL SYSTEM, THE MORE COMPLEX THE NOMINAL LEXICON CAN BE EXPECTED TO BE. As suggested by Maddieson (2005d: 55), there is some evidence that syllable structure complexity and consonant inventory size are interrelated cross-linguistically. This issue is discussed in more detail in § 5.4.2.8.

The correlations on a global scale already at this point help to explain some variation in particular areas of the world. For instance, it is common knowledge that there are significant differences in the size of consonant inventories in North America, to the effect that languages in the western part typically have larger inventories when compared with those in the east (Sherzer 1973: 774, Mithun 1999: 15).[7] Using the Rocky Mountains as a watershed dividing western from eastern languages, this difference turns out to be mirrored in the degree of morphologically complex lexical items as shown in figure 17, and

[6] Though note that variables such as length, nasalization and diphthongs are largely discarded in Maddieson's (2005h) coding scheme in order to make the data more readily comparable, and this approach is followed here for consistency.

[7] However, in North America, there is also "increasing head-marking as opposed to dependent marking going from west to east" (Fortescue 1998: 80). See § 5.4.2.12.5. for discussion of this as a possible factor.

this fact explains to some extent the areal hotspot of languages with a highly analyzable nominal lexicon detected in § 5.3.[8]

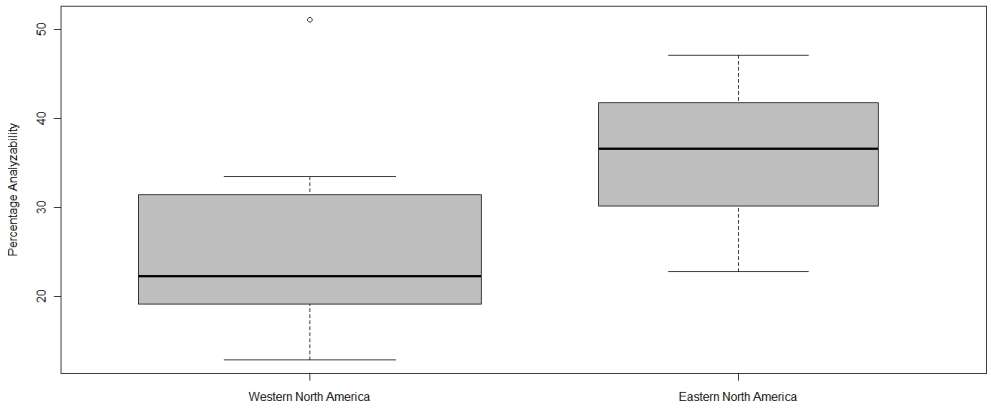


fig. 17: Differences in the percentage of analyzable terms between Western and Eastern North American languages

However, it is important to point out that these correlations are a statistical generalization, and there are languages which behave unexpectedly. In other words, there is no law in the sense of a classic implicational universal that a simple phonological system will in all cases trigger a lexicon characterized by morphological complexity. The most extreme case of a language that goes against the trend in the sample is Buin. Like its unrelated (or unrelatable) neighbor Rotokas, which is also spoken on the island of Bougainville, Buin has a very small phoneme inventory and simple (C)V syllable structure, and yet the degree of analyzable lexical items is one of the lowest both in the larger New Guinea area and worldwide (but see § 5.4.2.6. for a possible explanation of the behavior of Buin).

While establishment of a cross-linguistic correlation is of value in itself and in a way is more solid than proposed explanations for the correlation (Dryer 2003), it is important to note that skewed distributions are not an explanation in themselves, but rather something that needs to be explained (Cysouw 2003: 99), be it by appealing to functional, cognitive, or other factors. As a first step to get to the bottom of the correlations with phonological properties, and also in the light of concerns as uttered by Plank (2003: 138) that typology should not merely be an exercise in statistics, in the following sections three case studies will demonstrate apparent influences of phonological factors on complexity in the lexicon in synchrony and changes in diachrony in greater detail. Polynesian languages, Mandarin Chinese, and the "Papuan" language Vanimo will serve as examples. Particularly interesting is the case of Mandarin Chinese, for which there is actual diachronic evidence for the development of a largely compound-based lexicon and its phono-

[8] The outlier in Western North America is Kiliwa, which, although spoken in the west, has like other Yuman languages an average-sized consonant inventory as opposed to the large systems more common in the west.

logical motivations. For the case of Polynesian languages there are striking diachronic developments in the phonology that suggest a similar line of argumentation. Even though the earlier stages of Polynesian are not attested, they are at least fairly well reconstructed. The case studies will furthermore also serve to bring to light other aspects of phonological structure and its repercussions on the structure of the lexicon which can then be elaborated on.

5.4.2.3. Case Studies

5.4.2.3.1. *Case study I: Polynesian.* The Polynesian languages are a low-level branch of the Oceanic subfamily, which is in turn one of the best-established subgroups of the Malayo-Polynesian languages, themselves one of the primary branches of Austronesian. The Polynesian languages for which data were sampled (Hawaiian, Samoan, and Kapingamarangi) consistently score quite high with respect to the degree of analyzable terms, higher than most non-Polynesian Austronesian languages in the sample. At the same time, they are known to have very small phoneme, in particular consonant inventories. Since it is known that ancestral Proto-Oceanic had a considerably larger number of phonemes, and since the historical reconstruction of developments in that subgroup is in a fairly advanced stage, it should be possible to trace the developments in the Polynesian lexicon historically, departing from the Proto-Oceanic stage. Table 2 charts the Proto-Oceanic sound system as given by Ross (1988: 93).

	velarised bilabial	bilabial	alveolar	palatal	velar	postvelar
stop	bw	p b	t d	c j	k g	q
trill			r dr			
sibilant			s			
nasal	mw	m	n	ñ	ŋ	
liquid			l			ʀ
glide	w			y		

table 2: Proto-Oceanic consonant inventory, from Ross (1988: 93)

This amounts to a number of 23 consonant phonemes. Syllable structure was probably already fairly simple at this stage and is posited to be (C)V "with the option of a word-final consonant" (Lynch et al. 2002a: 66) which is lost in most daughter languages. Notably, loss of word-medial consonant clusters that were permitted in Proto-Malayo-Polynesian is one of the features that defines Proto-Oceanic as a subgroup (Lynch et al. 2002: 66).

Further down the genealogical tree, one finds the so-called Proto-Central-Pacific subgroup of Oceanic, believed to have been spoken between 100-800 BC on the Fiji islands (Trudgill 2004: 308, table 1, compiled from various sources; for a more general overview of the history of the Austronesian expansion see e.g. Pawley 1999). The Proto-Central-Pacific phase must have been relatively brief since evidence in terms of shared innovations is sparse, and it is thought to have been a dialect chain rather than a homogenous language (Pawley 1996b: 390; 2009: 529fn7). In any case, this dialect chain gave rise to both the Polynesian languages as well as Fijian and Rotuman, although the precise relationship of the latter to the other daughter languages is somewhat unclear. In other words, Proto-

Polynesian is a primary branch of Proto-Central-Pacific. The Proto-Central-Pacific consonant system (from Geraghty 1986: 290) is charted in table 3.

	bilabial	dental	alveolar liquids	alveolar fricatives	palatal	velar	labiovelar	glottal
fricatives	v			c	z	x		
stops	p	t	r			k	kw	ʔ
prenasalised obstruents	b	d	dr	s	j	q	qw	
nasals	m	n	l		ñ	g	gw	
glides	w				y			

table 3: Proto-Central-Pacific consonant inventory, from Geraghty (1986: 290).

Here one encounters 25 consonantal proto-phonemes, a little more than Proto-Oceanic had (Pawley, as cited in Trudgill 2004: 310, believes that the number of distinct segments was somewhat lower, "around 21"). Differences to the Proto-Oceanic situation include the presence of a contrastive series of prenasalized obstruents, a series of labiovelars and phonemic glottal stop.

Significant phonological simplification sets in on the way from Proto-Central-Pacific to Proto-Polynesian. Developments include (data from Geraghty 1986, see also Pawley 1996b: 392-393):

(i) merger of proto-phonemes */p/ and */b/, */d/ and */t/, */dr/ and */r/, */k/ and */q/, as well as */k/ and */kw/. In short, prenasalization is lost as a distinctive feature; these developments are shared with Rotuman. */r/ further apparently began to merge with */l/ in Proto-Polynesian under unclear conditions, a change that was completed in Proto-Nuclear-Polynesian. Proto-Fijian retains the majority of these contrasts.

(ii) Proto-Central-Polynesian */z/ changes to */h/ in Proto-Polynesian, in some instances the reflex is also *s*, i.e. a partical merger.

(iii) Merger of */j/ with */s/, */t/ or */d/; */j/ is only retained in Rotuman

(iv) Loss of */y/

(v) Merger of */ñ/ with */n/

(vi) Merger of labiovelars: */k/, */kw/, */q/, */qw/ fall together in */k/, and */g/ and */gw/ in */g/. Fijian retains the contrast between labiovelars and velars.

(vii) Merger of */x/ with */ʔ/

These developments leave Proto-Polynesian, most likely spoken between 500 BC and 200 AD on the Fiji islands (Trudgill 2004: 308, table 1), with a system of thirteen consonant phonemes (Biggs 1978): stops */p/, */t/, */k/ and */ʔ/, fricatives */f/, */s/ and */h/, nasals */m/, */n/ and */ŋ/, as well as */w/, */l/ and */r/.

Subsequently, Proto-Polynesian split into what is being called Proto-Nuclear-Polynesian, the common ancestor of the sampled languages Hawaiian, Samoan and

Kapingamarangi, on the one hand and Proto-Tongic on the other. As a result of this development, Proto-Nuclear Polynesian lost two further distinctive consonants inherited from Proto-Polynesian, namely */r/ and */h/, leaving it at eleven consonant phonemes. Proto-Central Eastern Polynesian, an even more direct ancestor of Hawaiian, additionally lost phonemic glottal stop, and, finally, Hawaiian itself is distinguished from its direct ancestor by merging nasals */n/ and */ŋ/ as well as */f/ and */h/, leading to its present-day system of eight consonant phonemes (glottal stop is reintroduced in Hawaiian by regular change of alveolar stops). The resulting Hawaiian phonological system allows for the generation of only 162 distinct syllables (Maddieson 1984: 22); it is seen in table 4.

	Bilabial	Dental-alveolar	Alveolar	Velar	Glottal
stops	p			k	ʔ
liquids			l		
fricatives					h
nasals	m	n			
glides	w				

table 4: Hawaiian consonant inventory, adapted from Biggs (1978: 708), Elbert (1979: 10-13)

Samoan also lost the inherited phonemic glottal stop just to reintroduce it as the regular reflex of */k/, but else maintains the Proto-Nuclear-Polynesian system, yielding the ten consonant phonemes seen in table 5.

	Bilabial	Labio-dental	Lamino-alveolar	Dorso-Palatal/ Dorso-Velar	Glottal
stops	p		t[9]	(k)	ʔ
liquids			(r), l		
fricatives		f, v	s		(h)
nasals	m		n	ŋ	
glides					

table 5: Samoan consonant inventory, adapted from Mosel and Hovdhaugen (1992: 20-21)

Samoan is characterized by pervasive diglossia. The above inventory is that of the tautala lelei; note that /k/ and /r/ are restricted to loanwords, and /h/ to loanwords and a few native interjections. The tautala leaga has three phonemes less, due to merger of /t/ and /k/, /n/ and /ŋ/, and /r/ and /l/. Mosel and Hovdhaugen also note that /p/ and /f/ are interchangeable for many speakers.

In Kapingamarangi, both */f/ and */s/ become /h/ by regular sound change, and the number of distinctive consonants in the language is thus nine, as seen in table 6.

[9] May "be pronounced as an apico-dental, apico-alveolar, lamino-dental, or lamino-alveolar stop." (Mosel and Hovdhaugen 1992: 20). /n/ may also be articulated as a lamino-dental or apico-alveolar (Mosel and Hovdhaugen 1992: 21).

	Bilabial	Dental	Alveolar	Velar	Glottal
stops	p	t		k	
liquids			l		
fricatives					h
nasals	m	n		ŋ	
glides	w				

table 6: Kapingamarangi consonant inventory, adapted from Biggs (1978: 708)[10]

Present-day Fijian, in contrast, has nineteen consonant phonemes (the Boumaa dialect in addition has phonemic glottal stop, Dixon 1988: 12). Table 7 shows the inventory.

	Bilabial	Labio-dental	Apico-dental	Apico-alveolar	Dorso-velar	Glottal
stops	p		t		k	ʔ
prenasalized stops	b		d		q	
liquids				r, dr, l		
fricatives	v	f	c	s		
affricates				j		
nasals	m		n		g	
glides				y	w	

table 7: Fijian consonant inventory (Boumaa dialect), adapted from Dixon (1988: 13)

In summary, "these unusually small inventories are simply the phonological end point of a millennia-long reduction in the number of consonants as languages spread further and further into the Pacific" (Trudgill 2004: 310).

Rensch (2002: 191) discusses these diachronic phonological developments, and states that, in connection with the simple syllable structure, "[t]he result is a high number of homonyms," which are fed in addition by syntagmatic phonological changes (for instance, to adduce an example from the present study, Hawaiian, Kapingamarangi, and Samoan *lā, laa*, and *lā* all colexify 'sun' with 'sail.' The terms were distinct in Proto-Polynesian, having the shape **laʻaa* and **laa* respectively according to Elbert and Pukui 1986: 188 and collapsed due to elision of intervocalic glottal stop). At the same time, Rensch relates these observations to statements in the literature as to "language inherent therapeutic devices which prevent or heal harmful clashes," to which he takes the Polynesian evidence to be a counterexample. However, it appears that an increase in segmental length of lexemes in Polynesian languages, and, on account of the evidence of this study, a substantial number of which by means of formation of complex lexemes, took place in

[10] Note that Lieber and Dikepa's (1974: 375) brief description of Kapingamarangi phonetics and phonology consistently distinguishes between slightly and heavily aspirated versions of each consonantal segment.

Polynesian languages, and this can be construed as just such a therapeutic device, a line of thought that will be discussed in much greater detail in following sections.[11]

In fact, the difference in morphologically complex lexical items in Polynesian when compared to the remaining Austronesian languages in the sample is consistently higher (figure 18).

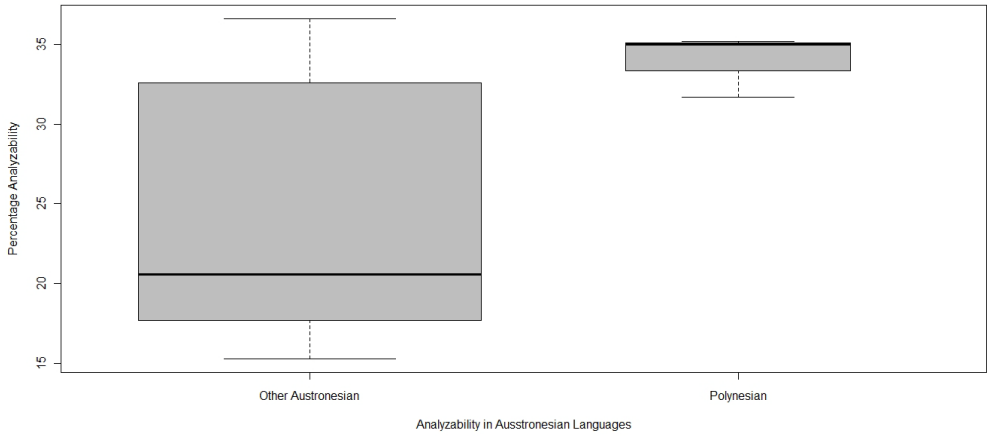


fig. 18: differences in the percentage of analyzable terms between the sampled Polynesian and other Austronesian languages

As becomes clear from the plot in figure 18, there is wide variation in the non-Polynesian languages, and some of the languages score as high as Polynesian. Notably, however, for those languages with high scores outside the Polynesian subgroup, similar accounts in terms of phonology are available: Tetun, for instance, independently has developed a small inventory of consonants phonemes (thirteen, according to van Engelenhoven and Williams-van Klinken 2005: 737, table 26.1) and concomitantly a lexicon that is characterized by a high degree of analyzability.

However, there are some unexpected results with respect to Austronesian that appear to run counter to the proposed account for the relatively high degree of analyzability in the lexicon of Polynesian languages. Fijian in fact does not do what one would

[11] Trudgill (2004), with reference to Polynesian specifically, argues that severe simplification of phonological systems is more tolerable in tightly-knit, isolated communities due to a high amount of shared information and cultural knowledge that can be pressuposed (while also noting that phonological systems of languages spoken by societies with the above characteristics may alternatively also be unusually large). Thus, in the case of Polynesian specifically, these societal factors, together with the absence of language contact with concomitant second-language acquisition did not prevent the development of very small phoneme inventories. In response to Trudgill, Pericliev (2004) tests the hypothesis of size of speech community and size of phoneme inventory on a large scale, with negative results, and Hajek (2004) argues that, at least in languages of New Guinea and the Pacific area, areal diffusion is apparently the most prominent factor responsible for the reduction of phonological inventories. For the purpose of the correlation between simple phonological systems and morphological complexity in the lexicon, one can remain agnostic as to what societal factors, if any, caused the significant decrease in inventory size in Polynesian, and simply note that this reduction did happen.

expect under the hypothesis. While it has a similarly simple syllable structure as Polynesian and a similar basic five-vowel system, its consonant inventory is far from being as drastically shrunk as the one of its Polynesian kin, but still the language shows a relatively high degree of analyzability in the lexicon that is comparable to that of Polynesian languages. Given its comparably large system of consonants, one would expect homonymy to be less of a problem in this language; however, Dixon (1988: 237), writing on the Boumaa dialect of Fijian specifically, informs that "[t]here is a good deal of homonymy in Fijian," and if this is indeed the case for whatever reason (e.g. a low functional load of some phonemes), then the same explanation of the structure of the lexicon is available for this case as well, in particular because the facts concerning syllable structure and vowel system fit the overall picture. Conversely, Rotuman, also a close congenitor of Polynesian, receives a low score in analyzability. It is at present unclear whether this is due to the phonemic inventory not being reduced as drastically as in Polynesian languages (with fourteen consonants, see table 8) or due to multiple layers of loanwords from a wide variety of sources (Biggs 1965, Schmidt 2003).

	Labial	Dental/Alveolar	Palatal	Velar	Glottal
stops	p	t		k	ʔ
liquids		r, l			
Affricate			tʃ		
fricatives	f, v	s			h
nasals	m	n		ŋ	
glides					

table 8: Rotuman consonant inventory, adapted from Vamarasi (2002: 7)

At present a convincing account of these facts is missing. As noted above, the data make a statistically significant cross-linguistic generalization possible, but it is far from being an absolute universal on a global scale, so counterexamples are neither unexpected nor damaging to the overall correlation. Thus, in spite of Fijian and Rotuman not quite fitting into the picture with respect to reduction of the consonant inventory, the difference between Polynesian languages and the other Austronesian languages in the sample is clearly present, and this difference is accountable in the way outlined above, the somewhat problematic case of Fijian notwithstanding.

Pawley (2009) primarily investigates retention rates in a variety of Oceanic languages spoken on the Solomon islands on the basis of a list of 60 vocabulary items, but also reports (2009: 529fn7) very high retention rates of basic vocabulary in both Proto-Central-Pacific (60 out of 60) and Proto-Polynesian (54 out of 60), which are therefore quite conservative Oceanic languages in terms of vocabulary replacement. Thus, in a time span of approximately 1,000 years after the breakup of Proto-Oceanic (though see below for potential problems with dating), a very large percentage of vocabulary items is retained in Proto-Polynesian. In contrast, percentages of retained vocabulary computed for purposes of glottochronology, as those in Elbert (1953) and Biggs (1978), indicate that among themselves, Polynesian languages have on average about 50 per cent shared vocabulary (as a remainder, such statements pertain to "basic" vocabulary as defined by the Swadesh list or similar lists), the highest figure being 72% shared vocabulary between Tongan and East

Uvea and the lowest 33% between Tongan, Samoan, and Tahitian. Thus, after the breakup of the still lexically conservative Proto-Polynesian, vocabulary replacement appears to have accelerated to a certain degree, and this may be a possible effect of the creation of morphologically complex neologisms replacing inherited vocabulary. However, "all the Fijian languages and some Polynesian languages (especially Tongan)" are considered lexically conservative when compared with some other Oceanic languages (Pawley and Ross 1995: 61), which speaks against replacement of inherited vocabulary on a larger scale.

On the other hand, one could also construct an argument in favor of relatively rapid vocabulary replacement out of the available data. The purpose of Pawley (1996b) is to defend from the point of view of linguistics the traditional view that posits a pause of around 1,000 years in the settlement of Eastern Polynesia after the settlement of Western Polynesia against claims by Irwin (1992, and other publications), who instead argues for more or less continuous settlement without major breaks. The linguistic correlate of that time span is the development of Proto-Polynesian out of Proto-Central-Pacific, and Pre-Polynesian is a term adopted by Pawley (1996b) to refer to the time before the breakup of Proto-Polynesian. Relying on glottochronological dates, Pawley (1996b: 400) notes that "[t]o allow only 400-500 years for the Pre Polynesian period would be to suppose a rate of lexical change over this period probably unparalleled in the subsequent history of any of the 30 individual Polynesian languages." Certainly, Pawley's argumentation is stringent, and this is not the place to contest archaeological evidence; however, it seems worth noting that, at least for the development of the lexical profile of Polynesian, the marked decrease in distinctive consonants on the way from Proto-Central-Pacific to Proto-Polynesian, which is continued in the Polynesian daughter languages, but which was already very advanced at the time of the breakup of Proto-Polynesian, may have accelerated lexical change in an unusually fast manner. There is no evidence for any other language present in the Fiji-Polynesia area at the time of Proto-Central-Pacific, and thus no indication that such accelerated rates of lexical change could be contact-induced (Pawley 1996b: 395). Within Polynesian, Pawley (1996b: 399), evaluating the data from Biggs (1978), states that "[t]he apparently more innovative languages include Samoan, Tahitian, Kapingamarangi and Nukuoro," two of which figure in the present sample. This again is compatible with the hypothesis of an increased rate of vocabulary replacement by coinage of complex terms because of limited expressive possibilities resulting from the shrunk consonant inventory, although the differential rates of vocabulary replacement, under the present account, still beg for a conclusive explanation, given that all Polynesian languages have experienced severe phonological simplification. Of course, any statements about the degree of vocabulary retention on the one hand and the degree of morphological complexity on the other are a function of the meanings selected for investigation, and if apparently conflicting results emerge, this may be attributable to the difference in vocabulary items that are investigated.

Summing up, the Polynesian case study is not entirely conclusive, and there are loose threads emerging from it that cannot be woven together into a coherent and conclusive account here. But the statistical difference between the degree of analyzablity in Polynesian when compared with other Austronesian languages remains a fact, as does the

heavy phonological simplification these languages have undergone since the time of Proto-Oceanic.

5.4.2.3.2. *Case study II: Mandarin Chinese*. There is one case where a temporal coincidence between phonological simplification and an increase in morphologically complex lexemes (compounds, in this case) is well-established and where the developments have been traced historically in a variety of publications: Mandarin Chinese. Typically, however, this process is discussed in phonological terms as disyllabification of the Mandarin lexicon, but, as will be seen in the following discussion, word-formation plays a major role in bringing about this pervasive change. Still, it is necessary to carefully distinguish between the phonological and morphological facts in the discussion (Feng 1997).

The basic facts concerning the simplification of Chinese phonology are as follows (dates from Arcodia 2007 throughout unless attributed to another author): According to Feng (1998: 213), syllables of CCVCC structure were possible in Old Chinese (ca. 1200 BC – 300 AD). Arcodia (2007: 84) and Feng (1998: 224) also mention clusters of up to three segments in both onset and coda in Old Chinese as spoken around 1000 BC. In Middle Chinese, in contrast, the syllable structure was simplified to CV(C) around 800 AD, with the additional constraint that only a subset of the available consonants, three nasals and three stops, were allowed in coda position. In Mandarin, only nasals appear in the syllable coda, initially three, later only two (Lin 2001: 84). In addition, whereas in Middle Chinese 35 distinct consonants could be found in onset position, only 20 are allowed for in Mandarin, and voicing was lost as a distinctive feature in consonants (Shi 2002: 73). Furthermore, affixation was lost. For instance, Old Chinese suffix *-s gives rise to a suprasegmental feature (tone) in Middle Chinese (Haudricourt 1954). Summing up, cluster simplification, loss of affixation, and reduction of possible consonants in syllable coda occurred. "As a result of consonant-cluster simplification, the number of phonologically distinct syllables in the language decreased dramatically" (Feng 1997: 213).

With the phonological simplification ongoing, the process of disyllabification of the lexicon set in. While it is true that disyllables are attested already in Old Chinese, it is equally true that they were relatively rare and that their number increased exponentially only at a later point of time. Text counts performed by Shi (2002: 75) suggest that the process of disyllabification (using disyllabic verbs as examples) reached its peak in the period between the 5th and 12th century AD; text counts by Feng (1997: 219) suggest an earlier date, to the effect that the process of disyllabification was "undergoing relatively large scale development during and after the Han dynasty" (Packard 2000: 265), that is between the 2nd century BC and the 2nd century AD. What are the precise mechanisms to disyllabify the lexicon that can be detected? Shi (2002: 76) lists the following:

(i) suffixation (Shi 2002: 74 mentions the nominal suffixes *-zi*, *-er*, and *-tou*)
(ii) "monosyllabic words are juxtaposed with synonyms," i.e. the creation of semantically redundant complex lexemes
(iii) replacement of inherited monosyllabic words by new disyllabic ones

(iv) reduplication
(v) conventionalization of adjacent syntactic constituents in discourse as fixed expressions that enter the lexicon, cf. Feng (1997: 208-209).

These facts make clear that the process of disyllabification is largely brought about by standard mechanisms of word formation (with a broad definition of word formation as employed for present purposes that does not exclude syntactic mechanisms from this category as long as they serve to form fixed expressions that enter the lexicon). Feng (1997) provides a number of enlightening examples, comparing a Classical Chinese text by Mencius (born around 370 BC), with a later commentary on the same text by Zhao Qi, written around 200 AD, that is, in the time of the Han dynasty in which disyllabification is said to have set in.

(2.) a. Mencius
shengren qie you guo
sage-person also have mistake
'Even sages make mistakes'

b. Zhao Qi
shengren qie you miu-wu
sage-person also have false-mistake
'Even sages make mistakes' (Feng 1997: 205)

(3.) a. Mencius
Wang Liang tianxia zhi jian gong ye
Wang Liang world 's lousy artisan PRT
'Wang Liang is the lousiest artisan in the whole world.'

b. Zhao Qi
Wang Liang tianxia bi-jian zhi gong-shi ye
Wang Liang world clumsy-lousy 's artisan-artisan PRT
'Wang Liang is the lousiest artisan in the whole world.'
(Feng 1997: 214, slightly adapted)

There are competing accounts for the increase in the number of (morphologically complex) disyllables while at the same time the language underwent phonological simplification, most prominently the 'functional' and the 'phonological' (Packard 2000: 266). According to the functional account, as summarized by Packard (2000: 266), societal and economic growth and concomitant introduction of new ideas during the Han dynasty (which is likely to be the time in which disyllabification of the lexicon set in on a larger scale) led to an increased need to coin neologisms in order to fill the gaps in the lexicon as no words existed to designate them. Once the lexicon was saturated with newly coined compounds, phonological distinctions, under this account, were given up since they were no longer

needed to keep words distinct, a job that had been taken over by the increased word length due to compounding. Also, it is argued that the increase in compounds were created by the preference in Chinese tradition to have pairs of entities, a solution which Feng (1997: 219) finds "theoretically unattractive, and empirically problematic." In contrast, the so-called phonological hypothesis (note that there are terminological inconsistencies: what is being called the phonological hypothesis by others is called the functional hypothesis in Feng 1997) states that the developments occurred rather in the reverse order, and that the increase in disyllabic lexemes is a functional response to the reduced complexity of the phonological system. Thus the label 'phonological hypothesis' is somewhat misleading, since it is at its core functional as well, albeit language-internally. This explanation is mentioned frequently, and is most often evaluated positively (Packard 2000: 265-267, Shi 2002: 72-74, see also Li and Thompson 1981: 14, and further references in Shi 2002). In the words of Lin (2001: 10):

> The change that started out with syllable simplification did not stop at the production of homophones. Indeed, one should not normally expect one change in a language to have no further effect, as chain reactions are common in language evolution. In the case of Mandarin, it is at least partially due to the great number of homophones in the language that another significant historical development was effected – the disyllabification of words. Earlier, we mentioned that M[iddle] C[hinese] had predominantly single-syllable words. However, when the syllable simplification was producing a great number of homophones, the dialect had to make some adjustment to avoid ambiguity. One logical measure would be to enlarge the word in size, and that was exactly what happened. ... Disyllabification has not wiped out the monosyllabic homophones; it has merely demoted them from the level of the word to the level of the morpheme in the dialect (Lin 2001: 10)

In favor of this account, Shi (2002: 74) importantly points out that southern varieties of Chinese preserve more traits of the inherited phonological system of Old and Middle Chinese when compared with the northern ones (including Mandarin).[12] This correlates with the fact that often southern Chinese monosyllabic words correspond to disyllabic compounds (probably of the semantically redundant type, see below) in northern varieties. "A simple explanation is that [southern] Cantonese has more phonological devices to distinguish lexical forms and thus does not need as many disyllabic words" (Shi 2002: 74).

Packard (2000: 267), in discussing the merits and drawbacks of the two accounts, also favors the phonology-based account "because it involves two processes that remain operative in the modern language: the continued simplification of the Chinese phonological system ... and the continuation of 'compounding' as a way of forming new words." Feng (1997: 213), however, raises some doubts regarding this explanation since the functional load formerly carried by segmental phonemes was in part taken over by suprasegmental features. Instead, Feng argues that the development of compounding is due to disyllabic foot formation that was established in the time of the Han dynasty, and which is itself due to the loss of bimoraic feet already occuring in Old Chinese. The simplified syllables re-

[12] Also, Mandarin has one of the smallest numbers of tonal opposition of any of the varieties of Sinitic (Mian Yan 2006).

sulted in a decline of syllable weight, to the effect that one syllable alone could not form the minimal prosodic unit of the foot anymore (Feng 1997: 226). Under this account, then, the phonological process of disyllabification is initially in terms of its motivation independent of the increase in compounds at the morphological level. In other words, Feng's account is a more sophisticated version of the phonological explanation, since the prosodic structure was ultimately caused by simplified syllable structure, and the causing factor here, as well as in more traditional versions of the phonological account, is ultimately simplification in phonology. Although Feng explicitly argues against traditional phonological explanation, because of the problem he sees with counter-functional compounds, Packard (1997: 7) summarizes his position as being an "insightful adaptation" of the traditional phonological view.[13]

5.4.2.3.3. *Case study III: Vanimo, Papua New Guinea.* An intriguing case in the literature for a correlation between extreme phonological simplicity and complexity in the lexicon (with examples almost exclusively drawn from the nominal domain) is Vanimo, a New Guinea language of the Skou family as discussed by Ross (1980).[14]

The segmental phoneme inventory features eight vowels, all of which may occur nasalized and sometimes contrasting phonemically with their non-nasalized counterparts. There are thirteen consonants and nine allowed consonant clusters, of which two are doubtful; in addition, there are three phonemic tones. Syllable structure is (C)V, where C is a single consonant, or one of the abovementioned clusters. However, for Vanimo specifically, a very important additional factor that appears to constrain the structure of the lexicon even more heavily than the sheer phonological facts is that "[t]he syllable and the morpheme appear to be – or to have been until quite recently – coterminous." By multiplying 20 consonants and (secure) consonant clusters with 16 vowel qualities and three tones, Ross establishes "that the number of possible morphemes in Vanimo cannot exceed – or have exceeded – 960, an extraordinarily low number. Semantically these resources are in effect less, as each verb paradigm has five or six different morphemes" (Ross 1980: 101). Note that there appears to be a correlation of such a situation, in which the syllable and morpheme are coextensive, with the presence of tones, as seen in the discussion of Mandarin Chinese. Ross states that the lexical concomitant of the phonological simplicity is the "attribution of very wide meanings to some morphemes, and their combination of other morphemes which act as specifiers." In the nominal domain specifically, these combinatorics primarily result in noun-adjective and noun-noun compounds. Ross (1980: 102-105) provides ample examples for the operation of compounding to counter the scarce distinctiveness of the language's morphological resources. A selection of examples, with some adaptations to simplify accessibility, are in tables 9 and 10.

[13] Similarly but independently, Duanmu (1999) argues that metrical structure favors disyllabic words (which have been present to a smaller degree already in older stages of the language and have been introduced to some extent by newly coined neologisms) in syntactic non-head position, which also accounts for the frequent semantic redundancy of Mandarin compounds emphasized throughout by him.

[14] Note that Skou, the language of the eponymous family in the sample, is not part of the core sample due to insufficient data as defined in chapter 3.

Simplex: *paŋ* 'arm, wing, frond'	
Complex term	Modifier
paŋè 'arm'	*è* 'bone, long object'
dìŋpaŋ 'wing'	*dìŋ* 'bird'
èŋpaŋ 'coconut frond'	*èŋ* 'coconut'
ñéŋpaŋ 'snake'	*ñéŋ* 'octopus'
yípaŋ 'sago frond'	*yí* 'sago pudding, food'

table 9: Vanimo compounds based on *paŋ*, adapted from Ross (1980: 103)

Simplex: *boŋ* 'intangible substance'	
Complex term	Modifier
yaboŋ 'smell, odour'	*ya* 'thing'
téboŋ 'smoke'	*té* 'fire'
øboŋ 'dust'	*ø* 'ground, earth'
høboŋ 'fog'	*hø* '??'
èŋboŋ 'coconut milk'	*èŋ* 'coconut'

table 10: Vanimo compounds based on *boŋ*, adapted from Ross (1980: 103)

Interestingly, Ross (1980: 101) also notes that "[n]oun compounding of this kind appears to be an areal feature of the West Sepik coastal region;" it is at present not entirely clear whether, if this is indeed an areal feature, it is due to similar grammatical restrictions as found in Vanimo. While not spoken in that area, there is another New Guinea language in which a very similar situation in the nominal lexicon obtains, and for which the same explanation is available: Toaripi. For this and for the closely related Orokolo, Brown (1972: 157) notes: "Both T[oaripi] and O[rokolo] have many homonyms or near homonyms and it often becomes necessary to guard against confusion of meaning. A way of doing this employed by both [languages] is to use compound expressions in place of the simple nouns." Brown does not mention what may have given rise to the situation of exuberant homonymy, but it seems extremely likely that phonology is the responsible factor here as well. Toaripi has nine consonant and eight vowel phonemes (Brown 1972: 119-120), and the syllable structure that can be inferred from the aforementioned source appears to be maximally (C)V.

5.4.2.4. Tonality and Morphological Complexity

Preliminary tests based on the WALS data revealed no discernible interaction between the presence or absence and the nature of tonality with the measured degree of morphological complexity. However, two of the case studies, that of Mandarin Chinese and the Papuan language Vanimo, revealed that there is a potential connection between tonality and the structure of the lexicon, which is why data on tonality for the languages in the statistics sample was gathered as well.

When assessing whether there is an impact of tonality on analyzability in the nominal lexicon, the *p*-value for tone as a predictor when distinguishing between simple and complex tone systems reaches only a very weak borderline significance at $p = .1057$. As the plot in figure 19 shows, the clearest contrast is between tonal- and non-tonal lan-

guages in general, while the differences between languages with simple as opposed to complex tone systems is not dramatic. In fact, instead of a constant upward trend analogous to rising complexity in the tone system, the analyzability score drops as one moves from simple to complex tone systems, which casts some doubt on the effect of tonality on the degree of analyzability. On the other hand, when simplifying the distinction to a binary opposition of tonality vs. non-tonality, the difference turns out to be significant at $p = .0342$ (estimate: 5.611).

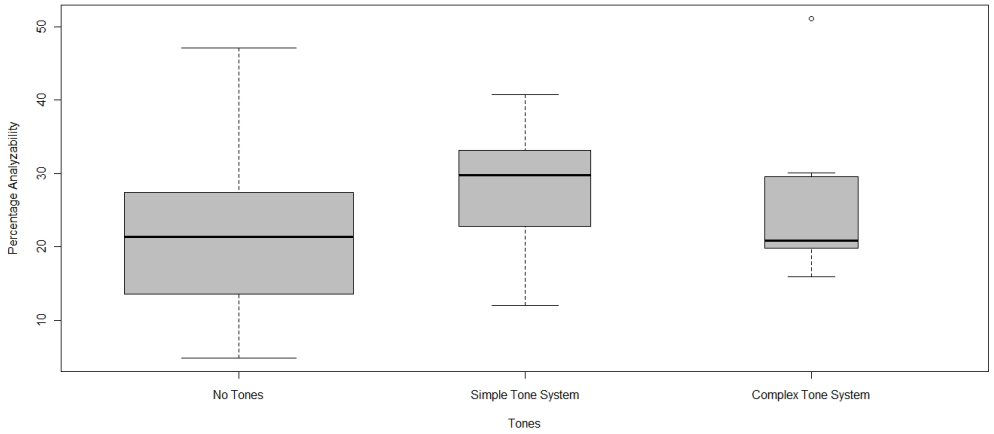


fig. 19: Correlation between tonal complexity and morphological complexity in the lexicon (data partly from Maddieson 2005e)

Thus, in spite of the unexpected non-linearity of the correlation, tonality can be added to the list of relevant phonological factors explaining morphological complexity: TONAL LANGUAGES TEND TO HAVE MORE ANALYZABLE NOMINAL LEXICONS THAN NON-TONAL ONES, an exception being Ket, a language with a relatively high degree of analyzability and thus unusual for Eurasia (see Vajda 2004b for an analysis that posits tones in Ket; however, the language is treated as non-tonal by Maddieson 2005e in the light of competing analyses).

On the one hand, the correlation between the presence of tones (whether the tone system is simple or complex) and an increased degree of analyzability in the lexicon is quite surprising in the light of Maddieson's (2005e) discussion of interrelations between tonality and other phonological properties. According to his data, increased tonal complexity typically goes hand in hand with a rise in the number of consonants as well as the number of distinctive vowel qualities (although he is also noting that the latter correlation in particular is subject to some areal variation, with a non-systematic relationship in particular in the Americas). In contrast, in the present study, one obtains a correlation between the degree of analyzability and tone as well as a correlation between this variable and consonant systems, which is quite surprising given that Maddieson's data indicate a correlation between tone and large segmental inventories! Maddieson (2005e) in particular suggests that a decrease in the complexity of the tone system goes hand in hand with a

decreasing number of languages with moderately complex syllable structure, while, as complexity in the tone system increases, the number of languages with complex syllable structure decreases.

On the other hand, the fact that there is a significant interaction between morphological complexity and tone is not too surprising, given that for instance the present-day tone system of Mandarin Chinese came into being as the complexity of syllables decreased. In fact, Mandarin Chinese is just one example for a broader scenario of tonogenesis outlined by Matisoff (1973). Without mentioning any particular reason why this should be so, Matisoff (1973: 77) states that as a prerequisite for the development of a full-fledged tone system "a language must have a basically *monosyllabic* structure (i.e. the morphemes must be only one syllable long)" and that "[t]here is something about the tightly structured nature of the syllable in monosyllabic languages which favors the shift in contrastive function from one phonological feature of the syllable to another" (Matisoff 1973: 28). This is in line with the observation made in the previous chapter as to the monosyllabicity of certain languages in Southeast Asia. Interestingly, one finds such a situation not only in the case study on Mandarin Chinese, but also in Vanimo and, in the Americas, for instance in Hupda (Epps 2008: 41 also notes "a strong preference for isomorphism between the morpheme and the syllable" in Hup, which has a two-way tonal contrast). According to this view, phonetic perturbations in the fundamental frequency of vowels due to neighboring consonants (see Hombert et al. 1979 for more phonetic details), which are an ordinary phonetic phenomenon, were phonemicized in Tibeto-Burman languages (which have a monosyllabic word structure) when phonological simplification broke down the originally complex phonological structure of the Tibeto-Burman monosyllables. In the words of Matisoff (1973: 79), "[i]t was only when the old consonantal system had decayed through cluster simplification, losses, mergers that the daughter languages were forced to exploit those pitch-differences for contrastive purposes." Importantly, in the highly abstract general scenario of tonogenesis as outlined by Matisoff (1973: 82-83), the impact of all these phonetic-phonological processes on the lexicon comes into play:

> Thus we may imagine a hypothetical language at Stage A: it is monosyllabic, but the number of possible syllables is very large, since there is a rich system of syllable initial and -final consonants. ... Different syllables have different pitches, but the language can afford to ignore this fact, since it is having no trouble keeping its utterances apart. [In stage B] its initial- and final-consonantal systems are breaking down. ... Homophony rears its ugly head. In desperation the language casts about for ways to protect its contrasts. Although each morpheme is still monosyllabic, the language now creates bisyllabic or even trisyllabic compounds in order to disambiguate homophones or near-homophones, so that the word is no longer monosyllabic. ... Meanwhile the number of vowels has increased and lexically contrastive tones have arisen, exploiting the previously redundant pitch-differences among syllables (emphases removed).

Matisoff (2001: 295) mentions that homophony is also notorious in the Loloish branch of Tibeto-Burman (compare also Bradley 2002: 1070), with compounding as a disambiguation strategy to counter it.[15]

Discussions of tonogenesis have a certain bias towards Southeast Asia, because the mechanism involved were first studied for languages of that area. However, there are also other possible diachronic paths leading to the emergence of phonemic tone. For instance, tonal contrasts in Cheyenne reflect Proto-Algonquian vowel length (Frantz 1972), with new lenght contrasts being introduced by the (sporadic, according to Goddard 1990: 104) loss of Proto-Algonquian *p and *k (Frantz 1972: 223). However, also in Cheyenne, the emergence of tonality goes hand in hand with at least some degree of segmental simplification, albeit of a different kind than for instance in Mandarin Chinese.

According to Ratliff (1992), a certain type of tone language (her Type A languages) can be defined by the fact that tone is used predominantly for contrastive lexical purposes, but only to a minor extent for morphological ones. Ratliff's example is White Hmong. This language has almost no segmental morphology, monosyllabic roots, a complex tone system, and a calculated number of 754 possible combinations of segmental contrasts without tonal contrasts factored in. According to Ratliff (1992: 135), "[s]ince syllables are usually coextensive with morphemes, almost all possible combinations need to be realized as morphemes. There is a high level of homophony as well," and thus, "[t]one must be used for lexical discrimination when there are not enough other resources avaliable in a tone language to do the job" (Ratliff 1992: 137). This statement is in agreement with Matisoff's diachronic scenario in which tone needs to be exploited to keep lexical morphemes distinct as phonological complexity decreases, next to an increase of the morphological complexity of words. Tone, as seen above, is suggested to be correlated cross-linguistically with monosyllabic words.

The correlations with phonological features are able to account for the behavior of many languages in the sample with respect to the degree of analyzability in their lexicon, but not all. For instance, Buin was already mentioned as an example of an "aberrant" language above. An entire region of the world where variation in analyzability cannot well be accounted for on the basis of the correlations so far established is the Caucasus. However, there is a way of accounting for this variation. This account is interrelated in a way with the relevance of the shape of the lexical morpheme for differences in analyzability suggested by the discussion of tone (although none of the Caucasian languages are usually described as being tonal, but see Kodzasov 1999, who argues that at least some Nakh-Daghestanian languages feature tone systems). Another reason to believe that this is a relevant factor comes from a number of languages with a relatively high degree of analyzability, tonal or non-tonal, for which authors note that lexical morphemes are normally monosyllabic, and that any elements departing from this shape in being longer can be identified diachronically as old compounds. This is the case for instance for Ket (Werner

[15] Matisoff (1973: 91n30) claims that "[i]nstances of this process abound in the world's languages. In some American English dialects where *pin* and *pen* are homophonous, the words are replaced by the compound forms 'stick-pin' /stɪkpɪn/ and 'ink-pen' /ŋkpɪn/, respectively."

1997: 46: "[h]istorisch lassen sich die meisten mehrsilbigen und auch manche einsilbige Wörter auf Komposita zurückführen" / "historically, most polysyllabic and also some monosyllabic words can be traced back to compounds") and Kiowa (Watkins 1984: 75: "there are polysyllabic nouns which can be tentatively regarded as old compounds on the basis of identification of at least one element with synchronically occurring forms. Still other polysyllabic nouns are entirely unanalyzable, but given the monosyllabic structure of roots and the tonal patterns of known compounds, they can safely be inferred to be old compounds").

But first, to make the argument more palpable and to show how it can account for variation that is otherwise not explainable, the following final case study presents the basic relevant facts about Caucasian languages.

5.4.2.5. Case Study iv: Variation in the Caucasus

There are three languages spoken in the Caucasus in the present sample, corresponding to the three major families that are indigenous in this region of the world: Abzakh Adyghe (Northwest Caucasian), Laz (Kartvelian), and Bezhta (Nakh-Daghestanian). These languages share a number of grammatical features, such as pervasiveness of ergative alignment. They also have some commonalities in the phonological systems, which typically feature a cross-linguistically unusual large number of consonant phonemes, to the effect that the Caucasus is sometimes said to form a linguistic area, although large-scale areality is disputable.[16] The languages are in addition all non-tonal (though again compare Kodzasov 1999 for a different point of view). Yet, there are also marked typological differences between the languages, and the Caucasus is also a region notable for its great linguistic diversity, both in terms of the large number of languages it hosts in a comparatively small territory as well as structural-typological variety (Comrie 2008). For instance, Northwest Caucasian and Kartvelian languages have many traits typically associated with polysynthesis, such as a rich system of verbal inflection. In contrast, morphological complexity is more pronounced in the inflection of nouns in Nakh-Daghestanian; particularly noteworthy are the rich case systems. There are sharp differences among the sampled Caucasian languages with respect to the degree of analyzable terms in the nominal lexicon that is presently investigated. Laz and Bezhta score very low and are thus typologically "normal" in the larger context of Eurasia, which is characterized by a comparatively low degree of analyzable terms when compared to the situation in the rest of the world (cf. § 5.3.). In contrast, Abzakh Adyghe is the language with the highest percentage of analyzable nouns in all of Eurasia.

The discussion in Rayfield (2002) makes clear that these differences can be accounted for by morphophonological factors. These factors, however, are less noticeable when examining the values assigned to the individual languages in the coding of their phonological properties. All are coded as having large consonant inventories and complex syllable structures. Bezhta and Laz have average-sized vowel inventories, while that of

[16] Tuite (1999), for instance, argues that the prevalence of ergativity in this region can equally well be explained by universal typological preferences, although not denying that the Caucasus has been a contact zone for a considerable amount of time.

Abzakh Adyghe is coded as being small. Rather, the structure of the lexicon, in particular the degree of analyzability, apparently has something to do in particular with restrictions on the phonotactic structure of the lexical root. As for the nominal lexicon of Kartvelian, according to Rayfield (2002: 1039), "the wide variety of syllable structures allow for a large number of non-homophonic roots, mono- and bi-syllabic" and the phonological inventories, together with the allowance for complex consonant clusters "give the language group enough resources to produce tens of thousands of distinct monosyllabic lexemes." Boeder (2005: 9-10) confirms the complexity of consonant clusters in Kartvelian languages, although noting that permittable clusters in Mingrelian and Laz are somewhat less complex than those of Georgian. Furthermore, in Kartvelian, there are marked differences with respect to phonological structure of the nominal and verbal root. "Nominal lexemes (and consequently denominative verbs) can show a complexity similar to Indo-European," while, in contrast, "[t]he core verb lexicon, depending heavily on a mono-consonantal root, is naturally characterized by frequent homophony" (Rayfield 2002: 1039; Gamkrelidze and Ivanov 1995: 768 also note that the canonical shape of root and affixal morphemes is identical in Kartvelian and Indo-European). Therefore, if the present study investigated the verbal domain, one could expect a rather different behavior of Kartvelian, and such differences in canonical structure between the verbal and the nominal root may well be partly responsible for the weak correlation between the values obtained for the present study and the overall analyzability of lexical items, including verbs, in the comparison with the World Loanword Database data in § 5.2.1.

The typical phonological structure of roots is very different in North-West Caucasian languages. There is little evidence for early contact with other Eurasian language families. This is in contrast to Kartvelian, which shows signs of early Indo-European influence or even co-evolution of lexical items. More importantly, as stated succinctly by Rayfield (2002: 1041), "Abkhaz and Circassian 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." The basic facts about Northwest Caucasian phonology and root structure are confirmed by statements of other scholars (among them Hewitt 2008: 307 and Nikolayev and Starostin 1994: 85, 192, who have it that the essentially monosyllabic root structure of Northwest Caucasian languages is due to loss of laryngeals and resonants from the more complex root structures in an earlier North Caucasian stage postulated by them), and is discussed for individual languages of the family. Kuipers (1960: 82-88) provides discussion of the situation in Kabardian, noting in particular the effects the canonical structure of lexical roots has on their semantics. Kuipers (1960: 87) discusses the example of the root *Ŝha* (written later on the same page as *Ŝẖa)*, which ranges semantically over "'head,' 'upper part' (roof, ceiling, summit, seed vessel of flower, ear of corn, riverhead), 'beginning' (of space, of time, crossing of roads), 'important part or member'

(place of honor, head of group), 'spherical part' (bulb), 'covering part' (sleeve), etc., also 'self'." Importantly, Kuipers (1960: 88) also points out that this situation is not much different from the semantic extensions of English *head*, but that still, "the two cases are by no means equivalent, as Kabardian lacks the numerous alternatives with a more limited semantic field that are found in English (roof, top, chief, bulb, etc.), so that polysemy plays a much larger role." Both in Kabardian as well as the sampled Abzakh Adyghe (Paris 1989: 161-162), there are combinations of consonants which act, from the point of view of phonology, as a single phoneme ("groupes consonantiques"). The lexical root in Abzakh Adyghe may consist of a single consonant or a consonantal group as defined above that can but need not be followed by a vowel, or of combinations of the two with insertion of epenthetic shwa (Paris 1989: 163). The apparent pronounced presence of homonymy in Northwest Caucasian languages is at first glance paradoxical, because the number of distinctive consonants is famously high. Thus Abzakh Adyghe only appears to go against the typological trend of having a large consonant inventory and a high degree of analyzability in the nominal lexicon. In fact, phonological restrictions on the level of the lexical root can be held accountable for its behavior. Rayfield's (2002: 1041) further discussion implies that this is less of a problem when it comes to the verbal domain, because the elaborate apparatus of affixation makes it possible to express semantic nuances that are not resolved by the "apparent lexical poverty" of the language, but for the domain of nominals lexical resources appear to be restricted (and note, interestingly, Rayfield's comparison with Chinese!). Concomitantly, Rayfield notes that "[t]he phonological structure of the language and, perhaps, a resistance to alien influences had led, where more sophisticated or abstract vocabulary is concerned, to fewer direct borrowings and more calques" (Hewitt 2005: 139, however, mentions cases of borrowing into Northwest Caucasian languages, but the proportion of borrowings may still be notably lower than in other Caucasian language families). Borrowing behavior is further discussed in § 5.4.2.7.1., but first a brief survey of the situation in the third language family of the Caucasus, Nakh-Daghestanian or Northeast Caucasian, is to follow. Rayfield (2002: 1041) characterizes the structure of the word and lexicon in this language family as assuming an intermediate position between Kartvelian and Northwest Caucasian. He notes, with special reference to the Nakh branch, that the permission of final consonant clusters and the frequency of di- and trisyllabic roots permit a reasonable number of distinct lexical items, while at the same time stating that especially the Chechen lexicon is characterized by a considerable number of homophones. One could thus assume from Rayfield's brief discussion that this intermediate position of the Nakh-Daghestanian family with respect to phonological restrictions on the lexical root inventory would lead Bezhta to have a degree of analyzable terms that is also intermediate between that of Kartvelian and Abzakh Adyghe. However, this is not so; the score for Bezhta is very similar to that of Laz. A quick browse through Comrie and Khalilov (2009a) reveals that most native noun roots in Bezhta have CVC or CVCV shape, which, given the very large inventories of both consonants and vowels,[17] allows for an ample amount of distinct monomorphemic lexical roots. Another reason for

[17] The former is typical of Nakh-Daghestanian languages, while the latter is unusually large compared with other closely related Tsezic languages.

the behavior of Bezhta in the context of the sample that comes to mind is a typological difference between Nakh-Daghestanian and the other two language families of the Caucasus: the former feature noun classes, and the noun class of the arguments are cross-referenced on the verb; this may provide a way to resolve lexical ambiguity on the discourse level (Rayfield's 2002: 1041 discussion also implies this scenario). However, casual inspection of the vocabularies of few or only one language is not sufficiently systematic evidence to show that root structure is a cross-linguistically operative factor. The following section attempts to explore this and related matters more systematically.

5.4.2.6. Canonical Structure of the nominal root

What the situation in the Caucasus shows is that, in general, it would be of great value for lexical typology to have a cross-linguistic study on the possible or typical phonological structure of basic non-derived lexical roots, both for the nominal and verbal domain. While reference grammars of course usually provide information on the syllable canon and phonotactic restrictions, information on the typical structure of lexical roots, and possible restrictions therein is less often found. It seems to be expectable that, if such information were more widely available, they would allow to show strong effects on the structure of the vocabulary, both in terms of the degree of analyzability of lexemes and possibly also on the degree of roots with a comparably vague and broad semantic content. In fact, it is plausible to assume that the interactions between phonology and the degree of analyzability would be further strengthened if this variable could be fully taken into account. As already mentioned, information on the canonical structure of lexical roots is not very frequently provided in reference grammars, but there are exceptions. For instance, Watkins (1974: 74) informs that the canonical shape of nominal (and verbal) roots in Kiowa is monosyllabic and of the shape (C)V(C), where certain consonants can also be followed by the palatal glide /y/, forming a cluster (Watkins 1974: 16), and the final consonant can only be /p,t,m,n,l,y/ (1974: 12-13). Furthermore, Conzemius (1929: 75) states that in Miskito, also a language with a nominal lexicon relatively rich in analyzable terms, "most words have been formed from a comparatively small number of elementary, monosyllabic roots." Miskito, in addition, has a small inventory of distinctive segments, and the basic morphological unit is in fact monosyllabic and the inventory of such units consequently severely limited by phonological factors, so one would expect morphologically complex terms to be relatively frequent in the lexicon, and this is exactly what is observable.

In order to assess canonical root structure in ideally all languages of the statistics sample, also when no such statements are found in the literature on them, the following interim procedure was applied: the number of syllables for all native lexical material in the database not coded as analyzable of any kind or semianalyzable were counted (with anything longer than four syllables, for ease of calculation, being counted as being tetrasyllabic), and then the weighted mean of the count was computed. This provides an empirical measure of the average length of the unanalyzable lexical morpheme in the language in question. However, a problem in obtaining reliable values is that the lexical data are at hand in orthographic, not phonological representation, and the challenge is thus to re-extract phonological structure, in particular syllabification, from orthography. A particularly problematic aspect of this are orthographic sequences of vowels, of which it is

not always clear whether they should be interpreted as diphthongs or sequences of vowels with a syllable break between them. Luckily, frequently such information is available, but there are eight languages, namely Mali, Toaripi, Kildin Saami, Cheyenne, Arabela, Cayapa, Chayahuita, and Cubeo, where vowel sequences in orthography are frequent and their interpretation remains unclear, and another one, Rotokas, where the source (Robinson 2011) briefly discusses the issue of syllabification of adjacent vowels, but remains non-committal as to the correct analysis. Since orthographic vowel sequences of up to five vowel graphemes are quite frequent for instance in Toaripi (Brown 1972: 132), any arbitrary decision as to their treatment engenders the danger of severely distorting the results, and thus, for this particular purpose, the abovementioned languages removed from the sample. In a number of other cases where the interpretation of vowel sequences is an issue, but where they are less pervasive, they were interpreted in a way that disfavors the hypothesis: in languages with a degree of analyzability lower than the cross-linguistic mean, where one would, by hypothesis, expect longer lexical items, they were treated as diphthongs, and in languages with a degree of analyzability higher than the cross-linguistic mean, they were interpreted as sequences. That is, if the procedure is biased in any way, it is biased slightly against the expected outcome.

Another issue is that, of course, the canonical structure of the native lexical morphemes are assessed only on the basis of a very small subset of all nominal items and thus may not be representative. However, for those languages where statements by experts are available, these are in very close agreement with the obtained weighted mean, so that the representativeness of the values seems granted. Table 11 provides these statements, together with the obtained weighted mean.

Language	Expert Statement	Obtained Weighted mean
Mbum	Hagege (1970: 63-64) reports that in his corpus, 55% of lexical items are monosyllabic, and 38% disyllabic.	1.808988764
Ket	"Ket basic vocabulary includes numerous non-derived stems, many of them monosyllabic." (Vajda 2004b: 14)	1.539473684
Carrier	"The primary roots are strictly monosyllabic, and they represent those objects or concepts, which are of the greatest import in American aboriginal life ...the Carrier language could be said to have some affinity to the monosyllabic idioms" (Morice 1932: 24) "In common with the primary roots, secondary roots express concepts or objects of simple import and are likewise unsynthetical substantives; but they are polysyllabic, generally disyllabic, in structure" (Morice 1932: 34).	1.426966292
Kiowa	"there are polysyllabic nouns which can be tentatively regarded as old compounds on the basis of identification of at least one element with synchronically occurring forms. Still other polysyllabic nouns are entirely unanalyzable, but given the monosyllabic structure of roots and the tonal patterns of known compounds, they can safely be inferred to be old compounds" (Watkins 1984: 75-76)	1.3

Itzaj	"Most noun roots are monosyllabic with the shapes CVC, CVVC, and CV'(V)C" (Hofling and Tesucún 2000: 87) "There are also polysyllabic noun roots of the form CVCVC or CVC(V)CVC ... Some of these are undoubtedly derived forms historically but are now considered to be unanalyzed forms." (Hofling and Tesucún 2000: 89)	1.138297872
Hupda	"While Hup strongly favors a syllable-morpheme isomorphism, it also permits words of more than one syllable; these, however, are almost all limited to two syllables. With the exception of ideophones ..., only a handful of words have three or more syllables." (Epps 2008: 80)	1.289855072
Jarawara	"...the language has a strong preference for roots with just two moras ..." (Dixon 2004: 71)	2.525641026
White Hmong	"Le hmong est une langue monosyllabique, les mots, pour l'immense majorité, n'étant formés que d'une syllabe." (Mottin 1978: 4)	1.014925373
Tetun	"Underived lexical morphemes in Tetun have from two to four syllables... most lexical morphemes are disyllabic." (Van Engelenhoven and Williams-van Klinken 2005: 739)	2.114285714

table 11: some expert statements for languages where they are available and computed weighted mean of the canonical structure of the lexical root

There is some areal variation and clusters of each type. For instance, in many languages of Southeast Asia, the canonical root structure is monosyllabic. The map in figure 20 shows the areal distribution of the types.

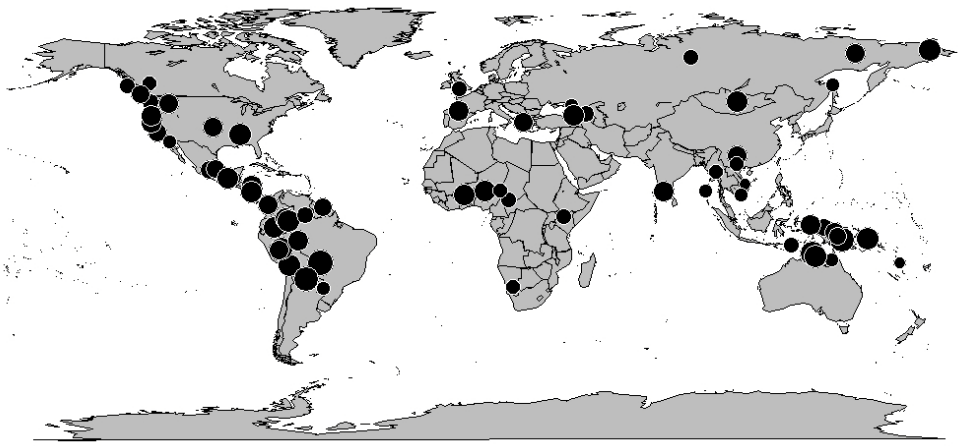


fig. 20: canonical length of the nominal root, reduced statistics sample

The same Mixed Model design employed already for systematic exploration of other phonological features was then used to analyze the data while controlling for area (canonical structure of the lexical stock is, like other phonological features such as those discussed above, susceptible to areal influence; one case in point are the Austronesian languages of the Chamic branch, which have adopted their inherited disyllabic roots to the common Southeast Asian monosyllabic structure, see e.g. Haudricourt 1956). As seen in the plot in figure 21, the same basic tendency already familiar from other phonological features can be observed: lower degrees of analyzability correlate with segmental complexity in nominal roots, and higher degrees of analyzability are found in languages in which the canonical root structure is more simple (the weighted means for each language were partitioned in four groups for visual representation only and are in Appendix C, but the actual more informative values themselves were used for statistical analysis).

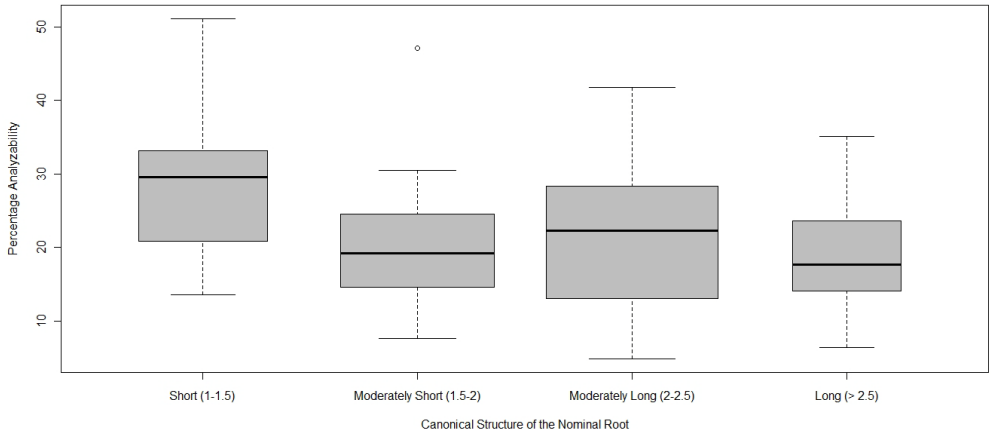


fig. 21: correlation between structure of the canonical nominal root and morphological complexity in the lexicon

Root structure does have a significant impact on the degree of analyzability in the nominal lexicon cross-linguistically when controlling for area (*p*-value for the predictor root structure: .0355, estimate: -5.111). Thus, bearing in mind the difficulties in the assessment of root structure and the hence somewhat error-prone methodology, THE NUMBER OF ANALYZABLE TERMS SEEMS INVERSELY CORRELATED WITH THE LENGTH OF THE CANONICAL ROOT IN MOST REGIONS. Given that this is the last of the features relating to complexity of the sound system and of the word to be discussed, it is now possible to convert the variable as to the type of analyzable lexical item (derived as opposed to lexical) into a cross-classificatory table, the other variable being the number of analyzable items, and add the typological correlate of complexity of the word and of the sound system.[18]

[18] Note that this table simplifies matters in that the degree of analyzabitility and percentage of derived vs. lexical terms are for ease of exposition treated as if these were absolute categories rather than the continua that they actually are.

	High degree of Analyzable Terms	Low Degree of Analyzable Terms
Lexical Dominating, Derived Subsidiary	• Low complexity in verbal person marking, fixed word order •Simple phonology, short roots	• Low complexity in verbal person marking, fixed word order •Complex phonology, long roots
Derived Dominating, Lexical Subsidiary	• High complexity in verbal person marking •Simple phonology, short roots	• High complexity in verbal person marking • Complex phonology, long roots

table 12: updated table showing the correlations obtained so far

5.4.2.7. Two Excursuses

5.4.2.7.1. *Excursus I: The linguistic treatment of items of acculturation, phonology, and overall complexity in the nominal lexicon.* In the discussion of the distribution of analyzability in the nominal lexicon in the Caucasus, it was noted that Abzakh Adyghe features relatively few loanwords when compared with the representatives of the other linguistic families of the Caucasus. Bezhta is rich in loanwords from Arabic, Avar, and more recently, Russian (see Comrie and Khalilov 2009b for full discussion), and Laz features many loans from Turkic, Greek, and Georgian. Further, as seen in § 5.2.2., analyzability in the semantic domains of both nature-related and body-part terms is strongly correlated with that in the domain of artifacts. Thus, one might be lead to hypothesize that the dominant technique a language employs to name novel artifacts, that is, whether it prefers borrowing or coinage of a neologism, is correlated with the degree of analyzable terms present in other areas of the lexicon: languages with many analyzable terms will typically more often accommodate items of acculturation by coining a neologism, while languages with a relatively high degree of simplex lexical items will more often respond by borrowing a name for novel objects from a contact language. Unfortunately, it is not possible to assess this prediction on a global scale on the basis of the sample. This is due to the fact that not all sources indicate the status of the listed lexical items, and it not advisable to attempt to identify loanwords by mere eyeballing, in particular because they are impossible to identify if one is unfamiliar with the donor language(s). Therefore, the discussion is restricted to languages of the Americas and to loanwords of European origin in the domain of artifacts, for two reasons: first, the sources consulted for this area of the world in the vast majority of cases indicate if a given lexical item is in fact borrowed, and should this be not the case, chances are high that loanwords can still be identified as such by inspecting their phonological shape since the donor languages are well-known European languages.

However, a certain margin of error obviously remains, and errors are possible. As elsewhere, it is possible that the same language features more than one term for the same concept, one of which may be borrowed and the other may be native but have experienced semantic extension or may be a morphologically complex neologism. In line with the policy in the overall assessment of morphological complexity, percentages are calcu-

lated, which is the reason why the global values reported in Appendix C are at times smaller than the number of loanwords listed.

Restricting the discussion to the Americas has another reason, namely that a differential degree of borrowing as opposed to coinage has been observed frequently here (e.g. Voegelin and Hymes 1953). The most comprehensive study on the topic is Brown (1999), who investigates the linguistic acculturation in languages of the Americas on the basis of a list for 73 items introduced by the Europeans. For each language in his large sample, Brown studies for how many of these items the languages have borrowed terms as opposed to other strategies of lexical expansion, and provides borrowing scores for each language. Since many languages of the present sample are also represented in that of Brown (1999), a direct comparison is often immediately possible. Relevant data are in Appendix C, where also further information that will become relevant for the present discussion is given: analyzability scores for all meanings except artifacts, as well as information for each language group as to which European power they were in contact with. Figure 22 plots the differential borrowing scores in the Americas obtained by this procedure; these will be used in the following analysis.

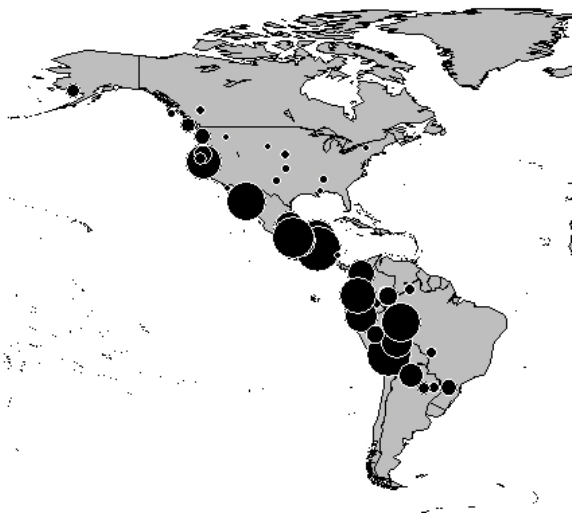


fig. 22: differential borrowing scores in the Americas

There is close agreement between Brown's and the scores obtained here. Although both studies sometimes employ the same source to extract the data, there is variation both in the number of items of acculturation and the individual items they investigate so the present study is not a mere replication of Brown's. What immediately strikes the eye in the map is the differential degree of borrowing depending on what the dominant contact language is. From Brown's data the generalization emerges that languages influenced by Russian, Spanish or Portuguese show a higher degree of borrowing than those influenced by English or French (Brown 1999: 80-81), with languages in contact with Spanish showing

the most pronouncedly high scores. In the words of Brown (1999: 81), "where direct Spanish influence has not been a factor, Amerindian languages have been disinclined to borrow European terms for items of acculturation." Taking up observations made by Bright (1960), Brown (1999: 81-82) relates this fact to the different ways in which Spanish as opposed to English- and French-speaking conquerors treated the indigenous populations which lead to different rates of bilingualism among Native Americans, and which in turn is thought by him to be responsible for the observed differential rates of borrowing. However, the details of the sociolinguistic situations are not elaborated on in depth by Brown.

Brown (1999: 83-91) also devotes space to discussion of the possible influence of structural features of languages on the rate of borrowing. Comparing variation in the degree of borrowing of genetically related languages, it emerges from Brown's study that sometimes languages from the same family, for instance Uto-Aztecan, show marked differences in the degree to which they adopted loanwords for items of acculturation. Like in this case, very often this degree of borrowing is correlated with what the contact language is, in line with the general observations made above: those Uto-Aztecan languages in contact with Spanish-speakers borrowed significantly more heavily than those in contact with English-speakers or French-speakers. For instance, Cora, which came in direct contact with Spanish-speakers, borrowed 80% of terms for the meanings investigated by Brown, whereas Comanche, which has been in direct contact with English- and French speakers and has undergone indirect influence from Spanish only, borrowed terms for only 17% (Brown 1999: 84, table 6.4.). Frequently, where there is little family-internal variation in the percentage of borrowed lexical items in Brown's study, as for Salishan, Siouan, Iroquoian, and Muskogean, it is the case that speakers of these languages had been uniformly exposed to contact with either the English and/or the French, which is further evidence for a scenario in which the dominant contact language is the major factor influencing the degree to which languages integrate loanwords for items of acculturation into their lexicon (by way of hypothesized different rates of bilingualism). Where there are significant differences in the number of loanwords in related languages that have been in contact with the same European languages, Brown tentatively resorts to language purism as an explanation (Brown 1999: 84-85).

A peculiar case is, however, that of the internal variance within the Yuman family. Kiliwa notably receives a loanword score of zero in both the present and in Brown's count, in spite of the contact language being Spanish. Mixco (1977: 20-21) explains the extreme paucity of loanwords to the difficult relations with and the hostility of the Kiliwa to Spanish culture. He also notes that other Yuman languages which are structurally similar to Kiliwa have borrowed more eagerly from Spanish and later also from English, giving figures of "approximately a hundred loanwords" in Diegueño and Paipai and fewer in other Yuman languages. Winter (1992) discusses the situation in Walapai, another Yuman language. Although noting that here there are a few loanwords from English and a somewhat larger number of loanwords from Spanish, Winter (1992: 219) says that "[i]t is widely assumed that Amerindian languages in general make wide use of descriptive terms, that is, of constructs whose parts taken together provide a composite reflection of crucial aspects of the meaning of the term." In Walapai, such morphologically complex terms are rather limited in native vocabulary, occurring most frequently in toponyms. However, in spite of

the comparably limited areas of application of complex terms in native vocabulary, "[i]t was precisely this technique which could be made use of to cope linguistically with a large influx of new notions from the culture of English-speaking Americans, short of taking over a great number of English words" (Winter 1992: 220). Winter's (1992: 222) summary is that the way the language dealt with acculturation was, in spite of a number of loanwords, "a strictly monolingual response in an increasingly bilingual situation."

Brown, in spite of arguing for bilingualism as the primary responsible factor for the differential degrees of loanwords in languages of the Americas, does not entirely rule out the possibility that structural features of languages may influence the degree to which they are eager to integrate loanwords, noting for instance the case of Salishan languages, which have accepted a larger number of loanwords than other North American languages not directly in contact with Spanish (Brown 1999: 90). However, he cautions that integration of lexical items of European origin into Salishan often was indirect via Chinook Jargon, and considers this explanation more plausible than one in terms of structural properties. In summary, Brown's (1999: 91) conclusion is that his data "suggest that if language structure factors affect lexical borrowing, they do so only minimally." That Brown attributes great importance to the contact language and the different sociohistorical circumstances of the contact scenario that come along is convincing, since these factors unmistakably are highly relevant. However, beneath these apparently major factors, there is some variation on a smaller scale that cannot be easily explained and that suggest that something else, even though probably subsidiary, is in play as well.

Though Salishan languages, according to Brown, have a relatively high loanword percentage when compared to other North American languages and this may be due to indirect borrowing via Chinook Jargon, it is still notable that languages spoken on the West Coast, such as Nuuchahnulth and Haida, although incorporating significantly less foreign lexical material than languages that underwent influence from Spanish, tend on average to also score higher on the loanword index than languages of Eastern North America. In § 5.4.2.2., a west-east cline of decreasing phonological complexity and concomitantly increasing analyzability of the lexicon was noted. Could it be the case that languages with a generally analyzable lexicon disfavor borrowing as the prime mechanism of lexical acculturation? This idea has been around at least since Sapir (1921/1970: 195-196), who suggests that resistance to borrowing has something to do with "the psychological attitude of the borrowing language itself." Comparing English and German, Sapir offers a psychologizing account of differences in the structure of the lexicon in terms of the lexicon and hypothesizes effects of these differences on the varying degree of borrowing in the two languages:

> English has long been striving for the completely unified, unanalyzed word, regardless of whether it is monosyllabic or polysyllabic. Such words as *credible, certitude, intangible* are entirely welcome in English because each represents a unitary, well-nuanced idea and because their formal analysis (*cred-ible, cert-itude, in-tang-ible*) is not a necessary act of the unconscious mind (*cred-, cert-,* and *tang-* have no real existence in English comparable to that of *good-* in *goodness*). A word like *intangible,* once it is acclimated, is nearly as simple a psychological entity as any radical monosyllable (say *vague, thin, grasp*). In German, however, polysyllabic words strive to analyze themselves into significant elements. Hence vast

> numbers of French and Latin words, borrowed at the height of certain cultural influences, could not maintain themselves in the language. Latin-German words like *kredibel* 'credible' and French-German words like *reussieren* 'to succeed' offered nothing that the unconscious mind could assimilate to its customary method of feeling and handling words.

Haugen (1956: 66) takes up this idea,[19] and Casagrande (1954: 228) suggests that, next to socio-historical factors, the paucity of loanwords in Comanche is attributable to the fact that "[w]ith an efficient means of word-building at hand, Comanche had little need to resort to linguistic borrowing."[20]

To test the hypothesis of a correlation between a general predilection for analyzability in native vocabulary and the relative degree of loanwords in languages of the Americas, values for the analyzability in the lexicon with the domain of artifacts removed in order to not replicate results were computed (the obtained values can be calculated from appendix B). A Generalized Linear Model, using both the degree of analyzability outside of the artifact domain and whether the dominant contact language is Spanish or Portuguese as opposed to English, French, or Russian was built.[21] Ineseño Chumash, Kashaya, Wappo, Carib, and Miskito, for which English and Spanish influence is about equally strong (though perhaps one of the languages was the dominant contact language at one time, and the other at another time), were removed from the calculation. To rule out possible effects from very closely related languages as well as spatial proximity and therefore potentially highly similar contact situations, only languages from different genera were subject to modeling. Note that this entails that the level of statistical independence is shifted down from the family to the genus level to still allow to include data from as many languages as possible for this particular test. Since the variable presently under investigation cannot be directly influenced by genetic inheritance, this seems appropriate for the present purpose. Data from languages not subject to modeling are presented in italics in appendix C. Modeling was begun by including an interaction factor between contact language and analyzability in native lexicon, which however appeared to be insignificant (p = .5097), suggesting that the parameters are independent or at least do not influence one another when it comes to the respective loanword percentages, and was hence removed from the model. The simpler overall model is highly significant (adjusted R^2 = .2942, $F_{2, 43}$ = 10.38, p = .0002099). As for the individual predictors, there was, unsurpris-

[19] Apparently independently, similar ideas are sketched by Ullmann (1962: 112-113), who also uses German as the example.

[20] Even Mixco (1965: 101) notes that "[n]ominal compounding is a productive syntactic process in Kiliwa" and that this fact "perhaps explains the paucity of Spanish loanwords." Thus, even for the case of Kiliwa, while sociocultural factors are probably the major force explaining the type of lexical acculturation dominant in the language, it may be aided by structural factors of the language.

[21] Note that this test operates with the assumption that loanwords are mostly found in items of acculturation, which in the slice of the vocabulary presently investigated clearly cluster in the domain of artifacts. This, however, does not rule out the possibility that languages also have borrowed from contact languages in other semantic domains, as is the case in some languages of Mesoamerica. Thus, there is the possibility that this fact skews the results in that loanwords, unlike calques, enter the lexicon of the borrowing language as unanalyzable wholes and may have replaced an analyzable native lexical item. A drawback of this approach is that it does not systematically control for this possibility.

ingly, a strong effect of the contact language (estimate: 25.9851, p = .000378), but notably also a weaker effect (estimate: - .9055) of the degree of analyzability that is also significant at p = .020730. Figure 23 plots the percentage of loanwords depending on the degree of analyzability in the remaining semantic domains investigated.

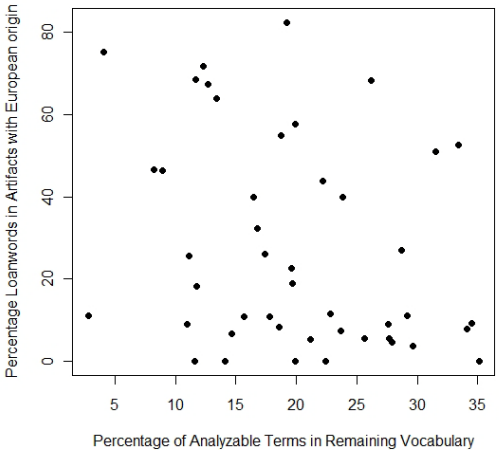


fig. 23: correlation between degree of borrowing and analyzability in the lexicon in languages of the Americas

The conclusion is as follows: IN THE AMERICAS, THE DEGREE OF BORROWING DEPENDS PREDOMINANTLY ON THE CONTACT LANGUAGE, BUT IS ALSO INVERSELY CORRELATED WITH THE DEGREE OF ANALYZABILITY IN THE LEXICON. This is in line with Sapir's statement: languages with an analyzable lexicon less readily accept loanwords than languages that have a larger number of monomorphemic lexical item. To reiterate, this statement should not be read as being equal to denying the overwhelming influence of which contact language is dominant and likely concomitant differences in bilingualism; but below the surface of this obvious difference, there does appear to be a more subtle influence of structural-organizational properties of the lexicon in general that does have an, albeit subordinate, effect on the degree to which a language is likely to accept borrowed terms for items of acculturation.[22] For the time being, the correlations that are obtained can only be said to be valid for the particular case study of the Americas, and it would be necessary to test in greater detail if this situation is demonstrable empirically also in other areas of the world.

[22] Sapir's (1921/1970: 195) position is in fact quite similar: He does not deny that the particular historical circumstances of the contact situation have to play a major role in accounting for differential rates of borrowing, but notes that "it is not the whole truth."

There is some evidence that there is a similar general world-wide trend from the data in the World Loanword Database (Haspelmath and Tadmor 2009c). Bradley Taylor (p.c.) kindly computed the simplicity score (as defined in Haspelmath and Tadmor 2009c) for the languages in the World Loanword database excluding loanwords (that is, those lexical items that are coded as clearly borrowed or probably borrowed). In effect, this score reflects the percentage of analyzable lexical items in native vocabulary (though as noted already above, this score takes into account complex items which are semantically redundant as well as semianalyzable terms). There is a certain trend for languages with relatively low ratios of morphological analyzability to have borrowed more lexical items than those with a higher degree of analyzability in native vocabulary on a global scale (Spearman's $\rho \approx .34$); however this positive correlation fails to reach statistical significance ($p \approx .11$).[23] It is plotted in figure 24.

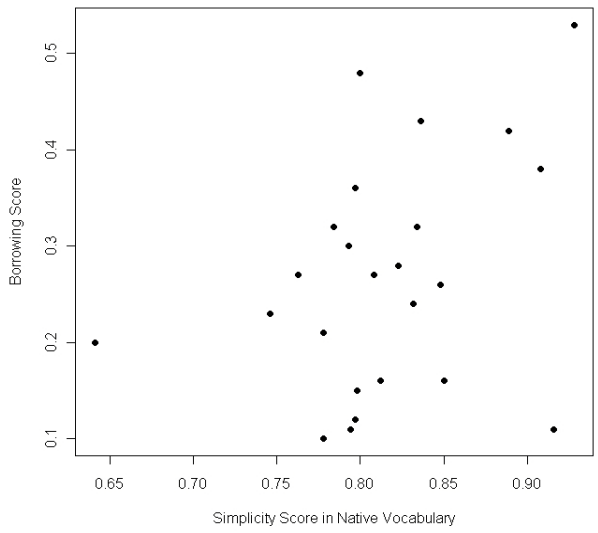


fig. 24: Simplicity Score in Native Vocabulary and Borrowing Score; data from the World Loanword Database (Haspelmath and Tadmor 2009c)

Thus the evidence from evaluation of the data from the World Loanword database is somewhat inconclusive, and it would require more in-depth research to either confirm or refute Sapir's (1921/1970) statement as to the influence of a relative paucity of morpho-

[23] Some of the languages in the World Loanword database are members of the same language family. To avoid possible biases from structural factors and to allow for statistical testing, one language per family was selected at random and the Creole languages Seychelles Creole and Saramaccan were excluded from analysis. The languages which entered calculation are, with their simplicity scores discarding loanwords following in parentheses: Bezhta (.784), Ceq Wong (.908), English (.889), Gurindji (.8), Hupda (Hup) (.797), Imbabura Quechua (.793), Japanese (.797), Carib (Kali'na) (.85), Kanuri (.778), Ket (.778), Kildin Saami (.823), Mandarin Chinese (.641), Mapudungun (.832), Oroqen (.916), Santiago Mexquititlan Otomi (Otomí) (.794), Q'eqchi' (.798), Swahili (.795), Takia (.834), Berber (.928), Thai (.848), Vietnamese (.808), White Hmong (.746), Wichí (.812), and Yaqui (.763).

logically complex terms in the lexicon and the predilection for accepting loanwords on a global scale empirically. In the above discussion, this was done by comparing the degree of analyzable terms in a selection of native vocabulary items. Indeed, one way to assess the productivity of a particular word-formation device suggested by Plag (1999) is to simply measure how many lexical items were created by its application. However, it is just one way, and since accounts such as Casagrande's (1954) explicitly refer to the productivity of the word-formation apparatus, rather than to the degree of analyzability in the conventionalized lexicon, a worthwhile investigation would be to also assess morphological productivity in other ways (see Plag 1999 for a number of suggestions as to how to measure productivity of derivational affixes in English). At any rate, a quite tentatively further correlate related to the degree of analyzable terms can be added, namely the differential rates to which the languages rely on borrowing as opposed to coinage of complex neologisms from the native stock of lexical items.

	High degree of Analyzable Terms	Low Degree of Analyzable Terms
Lexical Dominating, Derived Subsidiary	• Low complexity in verbal person marking, fixed word order	• Low complexity in verbal person marking, fixed word order
	•Simple phonology, short roots	•Complex phonology, long roots
	•Tentatively: favors neologisms	•Tentatively: favors borrowing
Derived Dominating, Lexical Subsidiary	• High complexity in verbal person marking	• High complexity in verbal person marking
	• Simple phonology, short roots	• Complex phonology, long roots
	•Tentatively: favors neologisms	•Tentatively: favors borrowing

table 13: updated table showing the correlations obtained so far

However, it must be emphasized that this result is tentative only and awaits further in-depth investigation.[24]

5.4.2.7.2. *Excursus II: A note on analyzability in Proto-Indo-European and other Eurasian reconstructs.* In Proto-Indo-European, the reconstructed ancestral language of the Indo-European languages, the canonical structure of the lexical root is monosyllabic (Szemerényi 1990: 130). The canonical Indo-European root is of CVC, or better CeC structure, with the consonant qualities being fixed and therefore root-defining and the vowel quality subject to systematic ablaut. The root can be augmented by resonants to yield structures such as *CReC, *CeRC, *CReRC, with **i* and **u* being capable of acting as resonants. In

[24] Rice (2012: 70-71) for instance, rejects internal structural factor as the cause of the high degree of motivation in the Athapascan language Dene Sųłiné, instead, inspired by Thurston (1989), arguing for little bilingualism as the more likely relevant factor (compare also § 5.4.2.12.1).

addition, there are constraints on possible root structure: roots cannot contain two plain voiced stops, or a voiceless stop and a voiced aspirate (Fortson 2004: 72, Szemerényi 1990: 99). As established by Benveniste (1935), the Indo-European root may be further expanded by a consonantal element ("root determinatives," "root extensions," "root enlargements") to form a stem of either verbal or nominal nature, yielding the form CVCC called 'theme I' and, for verbs only, CCVC called 'theme II.' Thus the root **pet-* 'fly' with the suffix *-er* yields **pét-r-*, continued in Sanskrit *pátra-*, and **pt-ér-*, continued in Greek *pterón* (Szemerényi 1990: 131); in both cases the meaning of the stem is 'wing.' Since roots are sometimes augmented by a preceding **s-*, of which it is sometimes unclear what governs its presence or absence (the so called s-mobile), the monosyllabic lexical morpheme of PIE can actually become complex with up to five consonants, with a CCCVCC structure, but "[e]ven apart from these reduced forms obtained by removal of the root determinatives, it can be empirically established that the majority of the monosyllabic roots contain only two consonants with the basic vowel *e* between them" (Szemerényi 1990: 131) and that "the structure of most PIE roots can be boiled down to a single template, **CeC-* (Fortson 2004: 70; CeC is, however, only the canonical root structure, and a number of roots with *a*-vocalism as well as non-canonical shape are found, Szemerényi 1990: 132, Fortson 2004: 72). As Lass (1994) points out with respect to the various extensions which need to be posited by the evidence from the daughter languages as augmenting canonically shaped roots, extensions of the root could be viewed as the "detritus" of old word-formation devices, the precise function of which cannot be recovered, given that a reconstruction of the PIE situation which posits that the extensions are part of the roots and cannot be segmented is unparsimonious in that it posits numerous synonymous and partly homophonous roots. What is more, Iverson and Salmons (1992) suggest, partly on grounds of typological naturalness, that even CVC root structure in Proto-Indo-European reflects a relatively late stage in the development of the language, with the consonant in the coda originally augmenting a simpler CV-type syllable structure and fused with the root already in the stage of the language that posits canonical CVC structure.

Importantly, many of these basic underived lexical roots within the lexicon of Proto-Indo-European are verbal in nature, with the root determinatives serving to derive both nouns and verbs, and further enhancements "always produce noun stems" (Szemerényi 1990: 131).[25] The above examples of roots augmented by a determinative, **pét-r-* and **pt-ér* 'wing' from the root **pet-* 'fly' already provides the transition to this aspect of the PIE lexicon, since in the case of the reconstructs one is dealing with analyzable terms of the derived type, more precisely, derived from a verbal root. In fact, Wodtko et al. (2008: xvi) note that "[g]erade deverbal motivierte Nomina stellen einen umfangreichen Teil des gemeinsamen indogermanischen Wortschatzes dar, da das urindogermanische Lexikon in stärkerem Maße deskriptive Mittel verwendet zu haben scheint, als es in vielen modernen

[25] The situation is in general in marked contrast to the situation in the neighboring languages of the Uralic family, in which disyllabic verbs and nouns or noun-verbs abound, with inflection and derivation obtained by suffixation (Janhunen 2001: 209). Uralic will be dealt with briefly later.

indogermanischen Sprachen der Fall ist"[26] / "deverbally motivated nominals in particular constitute a substantial part of the common Indo-European leixcon, as the Proto-Indo-European lexicon seems to have used descriptive means to a larger extent than is the case in many modern Indo-European languages," and Nichols (2010: 47) therefore calls PIE a verb-based language. Given that deverbal nominalizations are semantically and morphologically dependent on a verb, Wodtko's (2005: 50-51) conclusion is that they play a marginal role in the lexicon as mere makeshift devices that can when required be coined ad hoc, need not be learned, and are easily understood by way of being related to a verbal root. The important question as to the degree of conventionalization of deverbal nominalization which obviously cannot be answered for a reconstructed language put aside, this is a matter of the point of view one takes: if they are indeed frequent, then it could also be said that they, or rather the mechanism of nominalization per se, plays a major role in the organization of the PIE lexicon.

Unanalyzable nouns (that are by virtue of this of course also not deverbal) are, however, clearly also reconstructible for PIE. One type of athematic root nouns includes terms for "core vocabulary" meanings such as *h_3ekw- 'eye,' *ped- 'foot' and *dem- 'house' that probably represent an old stratum of the lexicon (though note that Rix and Kümmel 2001: 297, 458, as cited in Wodtko 2005: 63, posit verbal origins even for the terms for 'eye' and 'foot'). Like athematic nouns, there are also instances of nouns in the other major class of Indo-European nominals, thematic nouns, that are not relatable to other roots, among them generic level terms for animals and kinship terms such as *$u̯lk^wos$ 'wolf,' *$h_2ŕ̥tḱos$ 'bear,' *$snusós$ 'daughter-in-law' and *ag^wnos 'lamb' (Fortson 2004: 116; Wodtko 2005: 70-72 also mentions body-part, kinship and fauna terminology as the semantic domains in which monomorphemic nouns in PIE are found, see also § 5.4.1. for typological comparison). However, the majority of thematic nouns stand in a derivational relationship to known roots (Fortson 2004: 116, see Fortson 2004: 116-118 for an overview of noun-deriving processes). Wodtko et al. (2008: xiv) note that also the PIE root is capable of acting as a free-standing form, but still the root nouns ("Wurzelnomen") can be seen as an abstract or agent nominalization of the corresponding verb, see also Fortson (2004: 108-109) for an overview. Another type of root noun forms agent or undergoer nouns from verbal roots (Fortson 2004: 109); it appears to be these that Wodtko et al. (2008) are talking about.

Indeed, there are many unanalyzable lexical items in modern daughter languages, including many in "basic" vocabulary that can through comparative historical work be traced back and linked to stems based on typical CeC roots. Further, if the reconstructions are accurate, many PIE vocabulary items for the meanings on the wordlist used for the present study were analyzable in PIE, more precisely deverbal derivatives. Some assorted examples include those in table 14.

[26] A footnote by Wodtko et al. (2008) refers to Seiler (1975), whose work and elaboration on the notion of "descriptivity" was discussed in chapter 2. In fact, Seiler (1975: 38-39) briefly comments on the relationship between the frequently transparent relation between arguments and a predication by virtue of many nominals being derived from verbs and thus 'describing' their referent. He also suggests that this structure might be correlated with the absence or optionality of the copula in older Indo-European languages.

Root and gloss (original glosses in square brackets)	Derivative	Cognate of derivative (information in parentheses added)	Reference
*$b^herĝ^h$- 'become high, arise' ['hoch werden, sich erheben']	*$b^herĝ^h$-*o*-	Germanic **berga* (German *berg* 'mountain')	Wodtko et al. (2008: 30-31)
*h_1ed- 'bite, eat' ['beißen → essen']	*h_1d-*ont*-	Germanic **tanþ* (German *zahn* 'tooth')	Wodtko et al. (2008: 208, 210)
**sed*- 'sit down' ['sich setzen']	**ni-sd-ó*-	Old High German (and Modern German) *nest* 'nest'	Wodtko et al. (2008: 590-591)
*$h_2eḱ$- '(be/become/make) sharp, pointed' ['scharf, spitz (sein/werden/machen)']	*$h_2ák$-*mon*-	Lithuanian *akmuō* 'stone' ['Stein']	Wodtko et al. (2008: 287)[27]

table 14: examples of PIE deverbal derivatives

Of course, this is merely impressionistic and anecdotal evidence, and what would actually be required to allow for systematic exploration is a full 160-item wordlist for PIE, but the impression that the table above gives receives backup by experts on Indo-European as underscored by the quote from Wodtko et al. (2008) cited above, although one problem noted by Wodtko (2005: 52) are methodological difficulties in deciding whether a given derivative with reflexes in daughter languages does indeed entail that the derivative must be posited for the Proto-Language, since it could also be possible that the template for word-formation rather than the resulting form may have been inherited and daughter language terms coined independently on the basis of the common template.

Be that as it may, further questions that arise are: how natural is such a lexicon in which analyzability seems to be so pervasive cross-linguistically, and do other aspects of Proto-Indo-European as presently reconstructed accord with this observation to form a harmonic whole? With regard to semantics of the roots and lexical items derived from them, a lexicon as reconstructed by Pokorny (1959/1994), in which highly abstract meanings are dominant, are unnatural and implausible typologically (Sweetser 1990: 25-27), and these apparent shortcomings are likely due to the lack of a principled methodology of semantic reconstruction that does not generate a large number of highly abstract meanings for roots such as 'to swell' or 'to be bright' which abound in Pokorny (1959/1994) for reconstructs (a problem noted by Rix 2002: 1336). Put strongly, one could even say that a lexicon with such reconstructed semantics is a violation of the uniformitarian principle.[28]

But what about the sheer quantity of analyzability, regardless of the naturalness of semantic structure found in analyzable terms? Could a higher degree of analyzability in the Proto-Language also be an artifact of reconstruction, that is, does the very process of historical reconstruction of earlier stages of the lexicon of related languages necessarily

[27] Note also the PIE term for 'stone' mentioned in the very beginning of Chapter 1. This term is not mentioned by Wodtko et al. (2008).

[28] For instance, in the reconstructions proposed by Jóhannesson (1949) for PIE body-part terms, there is a conspicuously large number with a literal meaning of 'the curved one' or 'the swollen one.'

involve the discovery that synchronically unanalyzable lexemes in many cases can be traced back to morphologically complex ones? In a sense, this seems to be trivial, since complex terms are the norm rather than the exception for novel terms (Hagège 1993: 182-183 among many others), but is it a necessary concomitant of reconstruction, given the fact that after all one of the very task of etymological research is to make synchronically unanalyzable terms transparent by putting them in diachronic perspective (Rix 2002: 1336)?

At this point, the typological correlations established so far may help. In § 5.4.2.6., it was suggested that there is a correlation between the canonical structure of the lexical root with the degree of analyzability to the effect that the shorter the canonical root is, the more analyzable terms are found in the languages of the sample. Further phonological evidence is also available: assuming a standard non-glottalic reconstruction of the PIE consonant inventory with about 25 distinctive segments (15 stops in three series -voiceless, voiced, voiced aspirated- and five places of articulation -labial, dental, palatal, velar, labiovelar-, fricative *s, liquids *l and *r, nasals *m and *n, glides *i̯ and *u̯, and three laryngeals *h_1, *h_2, and *h_3, which would be an average-sized consonant inventory in terms of Maddieson 2005a), it becomes clear that the number of distinct roots with canonical shape this inventory is able to generate, not least due to the prevalence of *e*-vocalism, is clearly restricted; probably not as severely as the Vanimo system with the figure of 960 distinct morphemes calculated by Ross (1980), but also not unimaginably large (Jucquois 1966 counts about 2,000 attested roots from Pokorny 1959/1994). When it comes to the meanings expressable by these roots, the same is obviously true, and Jucquois (1966: 65, table 2) shows that the number of homophonous roots is very high, effectively reducing the number of 2,000 attested roots with distinct meanings to a much smaller number of attested roots with different phonological shape.

Thus, relating the evidence as to PIE root structure to the typological correlation between canonical root structure, size of the consonant inventory and analyzability in the lexicon, it is no surprise to find that the PIE lexicon appears to have been characterized by a high degree of analyzable terms. In general, leaving aside questions of details of reconstruction and the naturalness of the heavily root-based morphology of PIE, what the present study furthermore demonstrates is that A NOMINAL LEXICON THAT IS CHARACTERIZED BY ANALYZABILITY TO A DEGREE AS THAT APPARENTLY FOUND IN RECONSTRUCTED PIE IS NOT A TYPOLOGICAL ODDITY, which one might be inclined to think judging from the impression gained when comparing the reconstructed stage of PIE with modern daughter language or other better-known European languages, but has parallels in other languages of the world (see Comrie 1993 for discussion of the role of typological naturalness in historical reconstruction). As far as the aspects presently under investigation, a language like Kiliwa is typologically somewhat similar to Proto-Indo-European: an average-sized consonant inventory, with monosyllabic roots dominating the entire nominal and verbal lexicon, including a number of nonanalyzable nouns with this structure (see also § 5.4.1), but a large amount of nominals with more complex structure being either synchronically derived from verbs by a variety of morphological means or at least diachronically relatable to them (although the

nature of the derivational processes differs to some extent) and a high degree of analyzability in the lexicon in general.[29]

Another question that arises is: how did it come about that many modern daughter languages seem to be characterized by a markedly lower degree of analyzable terms when compared with their reconstructed progenitor? On the one hand, this observation hardly requires a special explanation, since it is an ubiquitous process for erstwhile morphologically complex terms to become phonologically reduced and demorphologized, in short, lexicalized as single unsegmentable wholes. On the other hand, there is evidence from at least two subbranches of Indo-European, Germanic and Slavic, that typological shifts took place in the lexicon which may have supported the transition from a largely analyzable nominal lexicon to a more unanalyzable one. Nichols (2009b) shows that in Slavic a lexical type shift from verb-based to noun-based took place. Kastovsky (2006a, b) demonstrates that in Germanic, there was a shift for the base form on which inflectional and morphological processes operate from the root-based type found in PIE to a stem-based type. This shift came into being by word-level stress becoming fixed in Germanic which made formerly predictable ablaut alternations unpredictable on the one hand, and on the other by an increase of secondary derived nouns and verbs. This ultimately led to the emergence of a new stem unit which served as the input of derivational processes (Kastovsky 2006b: 163). Still later, loss of medial and final unstressed syllables which were morphologically speaking markers of grammatical information occurred, ultimately leading to word-based morphology in Modern English. Of course, a logical concomitant of this development would be that inherited root-based derivatives would be reduced in their transparency in a perhaps more pervasive fashion than in the case of garden-variety lexicalization processes in individual items, as the productive apparatus of word-formation shifts to being based on the stem with derivational morphemes becoming reinterpreted as belonging to the stems.

Let us now turn to the question as to the potential artificiality of Proto-Language analyzability as a by-product of reconstruction, by comparing the reconstructed PIE state with that of two other Proto-languages of major language families of Eurasia, Uralic and Nakh-Daghestanian (in principle, it is not important that the language families are also located somewhere in Eurasia, and other language families might have been adduced as well). The reconstructed phoneme inventory of Proto-Uralic is somewhat smaller than that of PIE, with about 20 consonant phonemes (Rédei 1988: ix). As to the structure of the morpheme, demonstratives are reconstructed as monosyllabic and content words with very few exceptions as disyllabic, with the subtypes VCV, CVCV, VCCV, CVCCV, VCCCV, and CVCCCV (Rédei 1988: xi). Like Indo-European, Uralic is a deep family, with a primary split between the Finno-Ugric and Samoyedic subgroups. Janhunen (2009: 68) tentatively suggests a split of Proto-Uralic at 5,000 BP, but even the Finno-Ugric branch is assigned the proposed age of 4,500 years by Janhunen – plenty of time for diachronic change, in-

[29] One aspect not mentioned so far is that, as discussed for instance in Fortson (2004: 122-123), compounds are also reconstructible for PIE, and thus one could speculate that PIE belonged to languages of the mixed type as defined in chapter four for the present study; this would be another parallel to Kiliwa, in which both analyzable terms of the derived and lexical type are found.

cluding possible lexicalization of erstwhile morphologically complex lexical items to occur. Probing the Uralic lexicon for such processes at random using Rédei (1988) as a resource does indicate some phonological reduction and monosyllabification occurring in daughter languages when compared with the Proto-Uralic or Proto-Finno-Ugric reconstruct (one typical component of lexicalization). Importantly, however, there is no indication that in the reconstructed state of affairs, the parent term was morphologically analyzable, but rather appears to have been a unanalyzable word following the canonical disyllabic Uralic root structure. Table 16 illustrates this point, using the same set of four meanings listed above for Indo-European. In the case that states are reconstructible for several genealogical levels, the one for the highest level was selected as an example.

Meaning	Reconstruction	Cognates (selection, some marked as tentative by Redei); original glosses in square brackets	Reference
'mountain'	**kaδ'a*	Hungarian *hëgy* 'mountain, tip' ['Spitze; Berg'] Tas dialect of Selkup *ķée* 'hill' ['Hügel']	Redéi (1988: 115)
'tooth'	**piŋe* (Proto-Finno-Ugric level)	Finnish *pii* 'tooth, spike, peg, outer corner of house' ['Zahn, Zacke, Stift; äußere Hausecke'], Estonian *pii* 'spike, tooth, prong; sinew, muscle' ['Zacke, Zahn, Zinke; Sehne, Muskel'], Birsk dialect of Cheremis *püj*	Redéi (1988: 382)
'nest'	**pesä*	Finnish *pesä* 'nest,' Kildin and Notozero Saami *piess* 'Vogelnest,' Ezrä-Mordvin *pize*, Hungarian *fészëk* 'bird's nest; seat, abode' ['Vogelnest; Sitz, Wohnsitz]	Redéi (1988: 375)
'stone'	**pije*	Finnish *pii* 'firestone' ['Feuerstein'], Chantaika dialect Jenissej-Samojedic *fû*, Tawgy-Samojedic *fála*, Motor *hilä*	Redéi (1988: 378)

table 15: Some Proto-Uralic reconstructions and cognates in modern languages

While in this sense the report in table 15 is selective, it is not selective in that examples which do not involve erstwhile morphological complexity were deliberately chosen - the same situation is found in terms not listed in table 15, and thus there is no evidence that present-day simplex lexical items can be reduced on a larger scale to analyzable proto-language equivalents.

Nakh-Daghestanian is another ancient Eurasian language family. Nichols (2003: 297) considers the family to be at least 6,000 years old so that the age of this language family is also comparable to that of Indo-European (if not a bit older). The consonant inventory reconstructed for Proto-Nakh-Daghestanian is complex, and nominal root structure was canonically disyllabic, allowing for consonant clusters. Roots were required to contain at least one obstruent. In most languages of the Daghestanian branch initial consonant clusters are not allowed or can be shown to be secondary in some cases where they are found, but the reconstructed situation is preserved in Nakh languages (Nikolayev and

Starostin 1994: 82). Using data from Nikolayev and Starostin (1994)[30] for the same set of four meanings, the situation is in fact parallel to that found in Uralic, as seen in table 16: some phonological reduction in a number of daughter languages, but no evidence for erstwhile morphological complexity on the level of the proto-language.

Meaning	Root and Gloss	Cognates (selection)	Reference
'mountain'	**muɦialV* 'mountain'	Chechen *lam*, Avar *meʕér*, Archi *mul*	Nikolayev and Starostin (1994: 834)
'tooth'	**čɬɦV* 'tooth'[31]	Ingush *carg*, Bezhta *sila*, Tabasaran *slib*	Nikolayev and Starostin (1994: 326)
'nest'	**mōngwē* 'nest; bed'	Karata *minge*, Akushi Dialect of Dargwa *muga*, Lezgian *mug* 'nest, burrow; basket, hive; tree-hollow'	Nikolayev and Starostin (1994: 828)
'stone'	**hrŏmçwe* 'stone'	Botlikh *hinça,* Lak *nuwçi* 'iron or stone plate for roasting grain,' Khinalug *riçin*	Nikolayev and Starostin (1994: 495)

table 16: Some Proto-Nakh-Daghestanian reconstructions and cognates in modern languages

Again, the examples are meant primarily for illustrative purposes, and there are more synonymous or near-synonymous reconstructed lexical items not listed here which, however, also are not or do not appear to be morphologically complex. Notably, in addition, even for the rare cases where trysyllabic forms need to be reconstructed for Nakh-Daghestanian (such as **ʔVmx̱wĕlʔē* 'wild turkey,' Nikolayev and Starostin 1994: 225), there is no statement in the source that these are due to morphological complexity. Thus, there is no indication from the admittedly somewhat casual inspection of the reconstructed state of affairs that historical reconstruction necessarily leads to the establishment of Indo-European-style word families connected by a shared (verbal) root, and in this case, through this fact the reconstruction of PIE morphology, word structure and deverbal derivation gains plausibility precisely because it is not some inherent property of the method that is the cause for the reconstructed state of affairs.

If indeed PIE was a language characterized by a high degree of analyzability in the nominal lexicon as the evidence suggests, then this finding can be taken as an incentive to speculate about the behavior of the language in other related areas and thus to bring to

[30] Note that Nikolayev and Starostin (1994) entertain the controversial hypothesis that at an even deeper time depth Nakh-Daghestanian and Northwest Caucasian languages are genetically related ("North Caucasian"). For the present purpose, the search was restricted to lexical items reconstructible for the level of Proto-Nakh-Daghestanian to avoid making any commitment as to the accuracy of the claim of genetic relatedness between the two language groups.

[31] This root is reconstructed for Nikolayev and Starostin's "North Caucasian" level; it is offered here since no separate reconstruction for the Proto-Nakh-Daghestanian level is provided by them, although this should surely be possible.

light other aspects of the linguistic prehistory of Indo-European. One deverbal nominalization in PIE mentioned by Wodtko (2005: 61) is the word for 'plough,' *h_2arh_3-*tro*-*m*, consisting of the verbal root *h_2arh_3- 'to plough,' the instrument nominalizer *-*tro* and the nominative case suffix *-*m*. Wodtko further notes that a noun for this artifact that is independent of the verb was apparently not available to speakers of PIE, and adds in a footnote that this demand was also not met by borrowing from a contact language at the time of the Proto-Language. While this certainly is at first glance a trivial contingent fact about Indo-European, it is possible to actually ask the question: why not, and can this behavior be motivated? If indeed borrowing behavior should turn out to be related to analyzability in the native lexicon on a global scale, then one could expect the common ancestor of Indo-European languages to have had a dispreference for borrowing, but rather to have preferred coining neologisms (probably a considerable number of them by derivation from verbal roots) for artifacts using native lexical material. Thus, in this case, unfortunately for the task of establishing the areal context in which PIE was spoken by inferencing from loanwords, one could tentatively suggest that one should not expect to find much evidence for language contact as evidenced by apparent loanwords in PIE, since it would be natural for the language, given its typological characteristics, to have favored descriptive neologisms over loanwords. Leaving aside the vexing issue of the identification of the direction of borrowing, which can at times be even hard to determine in the case of actually spoken language, and even more so in the case of a reconstructed prehistoric language, there is some evidence for borrowed lexical material in the PIE lexicon. Gamkrelidze and Ivanov (1995: 769-776) even make a commitment as to the direction of borrowing, by stating that there are a number of loanwords from Semitic and Sumerian in PIE that predominantly denote domesticated animals and cultivated plants as well as names for particular tools and numerals, alongside loans from PIE in Kartvelian and languages of the Ancient Near East (some of the purported Semitic loanwords in PIE, such as the word for 'star,' are controversial however). If indeed there are loanwords in PIE, then this does not devalidate the hypothesis, which merely states that the number of loans should be relatively small. Note, however, that there are many ifs in the above statements; for one thing, the proposed account operates with the assumption that the situation found in the Americas is indeed replicable on a global scale which has not been demonstrated presently, and further, as already stated above, it would require a more systematic exploration of a larger portion of PIE vocabulary to consolidate the very fact that it was characterized by a large number of analyzable terms for a number of standardized meanings. Still, the case of PIE shows how typological data based on synchronic observations has the potential to contribute to questions of historical linguistics that are set quite removed from the present date, even though it can never be the only piece of evidence to solve puzzles of linguistic prehistory, which always requires detailed work by philologists.

5.4.2.8. Interactions between individual predictors

This section takes up the main thread of this chapter after the excursuses. Having established four apparently relevant phonological or morphophonological factors, namely size of the consonant inventory, complexity of the syllable, tonality, and length of nominal roots, it is important to assess whether these are independent of one another cross-

linguistically or linked in some way. In discussing the effect of differences in syllable structure complexity and consonant inventory size in § 5.4.2.2., Maddieson's (2005d) suggestion as to an interdependency between the two variables was pointed out. While in the words of Maddieson (2005h: 15), "absolutely no correlation was found between the number of vowels and the number of consonants" (see also Justeson and Stephens 1984 for a full-length study), Maddieson (2005d) notes that there is a correlation in his sample of 484 languages between the structure of the syllable and the consonant inventory size, to the effect that with increasing complexity in the syllable structure there is a rise in the mean of consonant inventory size in the languages of his sample, as seen in table 17.

Syllable Structure	Average Number of Consonants
Simple	19.1
Moderately Complex	22.0
Complex	25.8

table 17: Average number of consonants for languages with different levels of complexity in syllable structure, adapted from Maddieson (2005d)

However, as acknowledged by Maddieson (2005d: 55) himself, his sample is neither controlled for genetic nor areal effects, and thus he cautions that the results may be due to fortitous historical contingencies rather than a genuine "design feature of language." Furthermore, in the discussion of the diachrony of tone, a correlation between tonality and complexity in the syllable structure as well as the structure of the lexical root were alluded to.

Given these different suggestions as to interactions between the variables that play a role in shaping the structure of the lexicon, it is imperative to test in a systematic fashion what correlations exist between the relevant variables concerning complexity in phonology and root structure in the present sample to obtain a better understanding as to which phonological features are really relevant in shaping the degree of analyzability in the nominal lexicon. When assessing dependencies of consonant inventories and syllable structure putatively identified by Maddieson (2005d) on the basis of the languages in the statistics sample, which has the property of being genetically balanced, indeed such a dependency is found (Spearman's $\rho = .3$, $p = .01276$; in this case, unfortunately no model also taking into account areal factors is possible because the residuals do not fulfil the required assumptions). The associated plot is in figure 25.

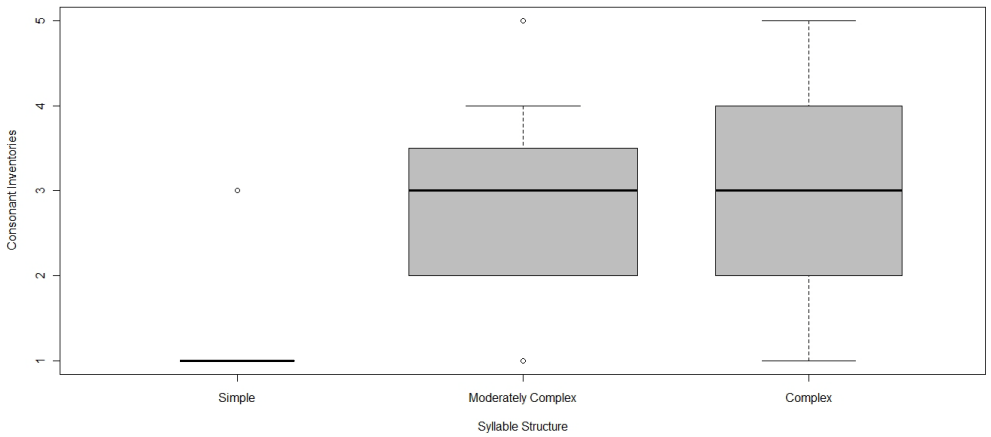


fig. 25: Size of consonant inventory correlated with complexity in syllable structure

Thus, at least speaking for the sampled languages, the two measures are not independent (and Maddieson's larger sample suggests that this might be also true on a larger scale). Now, this fact does not damage the findings regarding the effect of phonological factors on the lexicon: rather than acting as independent factors influencing the structure of the lexicon, one could then say that they "team up" and together exert influence on the degree of analyzable lexemes in the lexicon. But which factor, if any, is more important?

Moreover, as visualized in figure 26, canonical structure of the nominal root was found to be predictable (p = .022, estimate: -.1061) by the size of the consonant inventory: the larger the consonant inventory is, the shorter are the lexical roots. This correlation is similar to that found by Nettle (1995, 1998) based on smaller samples.

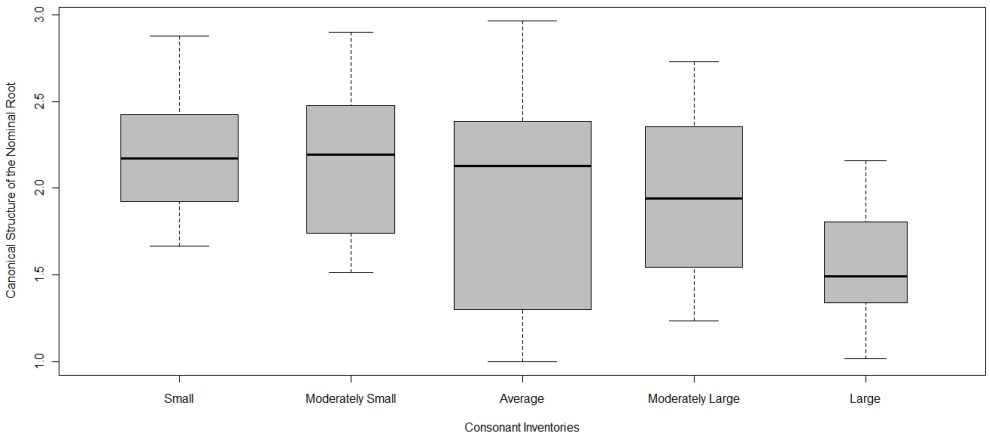


fig. 26: Size of consonant inventory correlated with canonical structure of the nominal root

Moreover, unsurprisingly given the suggestions in the literature, canonical root structure is predictable by tonality (p = .0014, estimates: -.3615 and -.6817). In languages with complex tone systems, length of nominal roots drops dramatically (not just in Southeast Asia), as seen in figure 27.

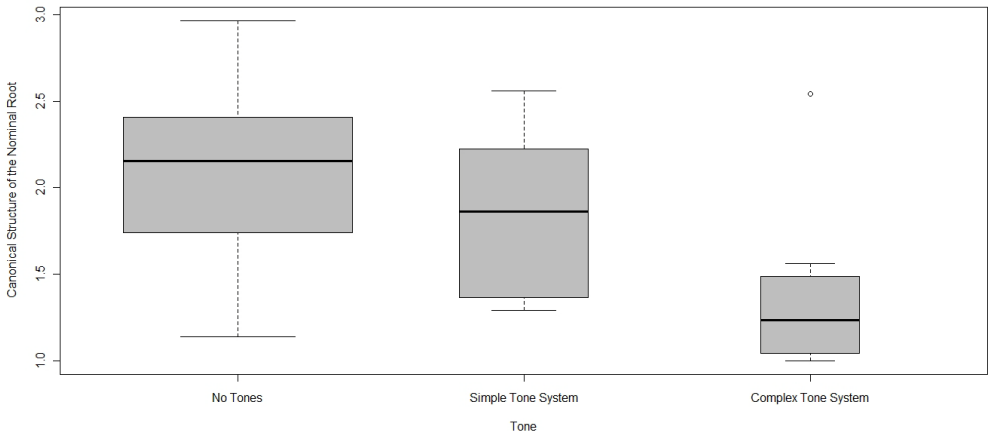


fig. 27: Canonical structure of the nominal root correlated with tonal complexity

Thus, summing up the evidence so far, there are four to some extent interrelated phonological factors interacting with the degree of analyzability in the nominal lexicon to a statistically significant degree: the size of the consonant inventory, which is itself correlated positively with complexity in syllable structure and negatively with the canonical shape of the nominal root. This latter factor in turn interacts with tone, to the effect that when tonal complexity increases, roots become shorter. Figure 28 summarizes the dependencies diagrammatically, with black arrows between features indicating a dependency.

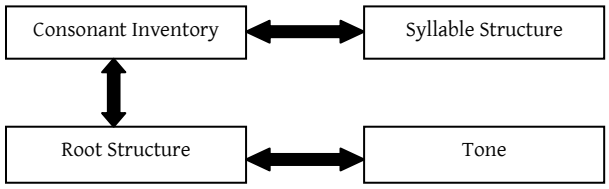


fig. 28: correlations between different aspects of morphophonological structure

Obvious and important questions that arise from these findings are thus (i) whether it can be assessed if one or a subset of the phonological properties is really the relevant one for the behavior of the sample languages with respect to analyzability, with the other(s) being a side effect due to interactions in phonology that are independent of this and (ii) how

precisely the features interact with each other, i.e. whether there are combinations of feature values that give rise to particularly high (or low) degrees of lexical analyzability.

To answer these questions, however, more complex statistical analyses would be called for, for instance a General Linearized Mixed Effects Model. While this is not in principle a problem, it is an issue because the sample size of the present study is relatively small, and thus not all logically possible combinations of values are attested in the sample. This already becomes a problem when trying to take into account only two features. For instance, there are no languages in the sample with complex tone systems and simple or moderately complex syllable structure, there are no languages in the sample with large consonant inventories and a simple syllable structure, and so on (and for many, but not all, other combinations of values there is just a single language in which it is realized). When combining all three relevant features, the coverage becomes even more fragmentary, and there are very many combinations of values which are simply not attested in the sample. This is a situation that is detrimental for the reliability of statistical analyses and the conclusions that can be drawn from them, because statistical power of the model is then low and it becomes instable in that very small changes in the data can have dramatic effects.

What the dependencies between the individual variables however at any rate do show is that there is every reason to believe that the features interact in significant ways, and that their effects combine in exerting influence on the structure of the lexicon. In the absence of reliable possibilities of statistical testing, this can be shown in the following fashion: when values for the segmental phonological variables that showed significant interaction in the lexicon are combined to a single index (bypassing tone, both because the decrease in analyzability as tone systems become complex is hard to interpret and because here the correlation is positive and thus hard to integrate into a combined measure with the otherwise consistently negative correlations), effects become very strong. Combining the individual variables is done by conflating the information they provide into one variable, which will be called the COMBINED PHONOLOGICAL COMPLEXITY INDEX (CPCI) in the following for want of a better term. The CPCI is computed in the following way: first, the value for the canonical shape of the native unanalyzable lexical morpheme is scaled down to ordinal scale with four levels of variation (as is done for the plot in figure 21). Languages with values between 1 and 1.5 are grouped together, and so are those with values between 1.5 and 2, 2 and 2.5, and 2.5 to 3. This entails that some information is lost, but the procedure is statistically valid nevertheless. Now, there are three variables: consonant inventories with five levels, syllable structure with three levels, and root structure with four levels. In order to normalize the different scales and thus to render the values comparable, they are multiplied to reach the smallest common denominator, which is 60. Thus, the value for consonant inventories is multiplied times twelve, that for syllable structure times twenty, and that for root structure times fifteen. These values are then added up and the sum is divided by the number of attested values, ideally three, but sometimes only two due to lack of secure data (if only one feature value is available, the CPCI is not calculated). Values for the CPCI are in Appendix C. In a Mixed Effect Models, with the percentage of analyzable terms as a response value, the CPCI as a fixed effect and area as a random effect, there is a very significant impact of the CPCI on the analyzability score at $p < .0001$. Figure 29 plots the results.

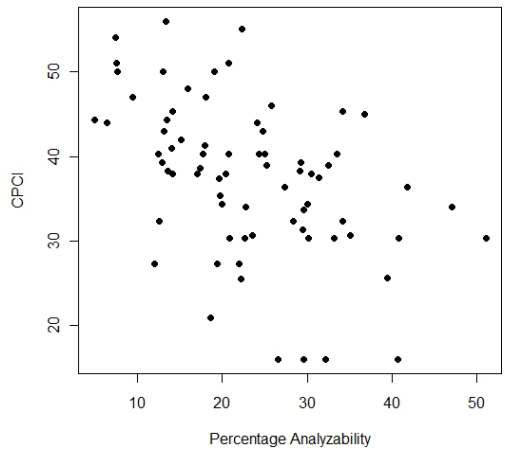


fig. 29: correlation between the combined phonological complexity index and analyzability

Thus, a combination of the individual features leads to a much stronger correlation with analyzability in the lexicon than when each is taken as a fixed effect on its own. And this is hardly surprising, given that languages such as Pawnee and Bororo, which have both simple syllable structure, small consonant inventories and relatively short lexical roots, are situated on the upper end of the continuum in the number of languages with analyzable terms.

5.4.2.9 Intra-family variation in phonological complexity and analyzability

It is also possible to ask whether the same principle that is operative typologically is also observable within language families (see Greenberg 1978, 1995 for intra- and intergenetic comparison respectively). Given that genealogically related languages started out from a common state with respect to the phonological system, complexification or simplification in that system would be expected to have an effect on analyzability in the lexicon. And if this prediction turns out to be true, it would be evidence for the operation of some sort of diachronic pressure that causes languages to adapt with respect to one of the variables as the other changes.

For this purpose, CPCIs were computed for all languages in the entire (EXT-2) sample which (i) fulfil the criterion of having more than 65% of equivalent terms for the meanings available for comparison and (ii) are known to go back to a common ancestral language. Criterion (ii) was applied rather strictly in this context, that is, the comparison was only carried out if the genetic relationship is firmly established and not controversial among experts of the families in question (for this reason, for instance, Kanuri was not compared with Dongolese Nubian and Ngambay since the genetic link as parts of Nilo-Saharan between them, in particular concerning Kanuri, is not uncontroversial; similarly, comparison was not carried out between languages classified as Pama-Nyungan in Dryer

2005a, although the general trend to be reported there is also observable here, and neither were "Hokan" languages compared). The obtained results are in Appendix C. What they show is that, for the eleven language families for which only two languages are compared, in seven, the language with the higher CPCI has the lower number in analyzable items, and that with the lower value in the phonological complexity index the higher degree of them. Data from three families, Afro-Asiatic, Jivaroan and Quechuan, run counter to this trend, while the evaluation for Nilo-Saharan is equivocal since both relevant languages receive the same score for the CPCI. Where more than two languages enter into the comparison, values are not always paired hierarchically, but here, correlation tests are available to assess the dependency between the two variables. This turns out to be always negative, as expected. It is relatively weak in the case of Niger-Congo (Spearman's ρ = -.08571429), and quite strong in Austronesian (Spearman's ρ = -.6107894), Sino-Tibetan (Spearman's ρ = -.5), and Uto-Aztecan (Spearman's ρ = -.5). Thus, THE SAME COVARIATION BETWEEN PHONOLOGICAL AND LEXICAL COMPLEXITY THAT IS OBSERVABLE IN AN INTER-FAMILY TYPOLOGICAL COMPARISON IS ALSO OFTEN NOTICEABLE WITHIN LANGUAGE FAMILIES THEMSELVES.

However, there is the issue that sometimes, the languages of the same family are quite heterogeneous typologically, and in addition, their sociolinguistic status may be vastly different (take, for instance, Manange vs. Mandarin; for the same reason, arguing on the basis of the Austronesian data is dispreferred due to differences in societal scale). While the hypothesis presently entertained is that phonological factors alone are dominantly responsible for the variation in analyzability, there may be other yet undetected interfering factors in the play (see also § 5.4.2.12), and these may cause the correlation to be altered in either direction. One may think of diverging grammatical organization, but also differences in sociolinguistic setting, including factors such as the number of speakers, the size of the territory they occupy, whether or not the language is learned as a second language, etc. Therefore, a particularly useful test case would be one in which both differences in grammatical as well as social structure are reducible to a minimum so that other factors are unlikely to play a big role, the only major difference between the languages of the same family lying in the phonological system. In other words, the variation should be confined exclusively to the variables in question, with everything else being as similar as possible (such an approach in comparative linguistics is first applied by Pederson 1993 and taken up by Bickel 2003). The languages in the sample that come closest to this ideal situation are Cayapa and Tsafiki, both members of the small (five languages) Barbacoan language family of lowland Ecuador and Columbia. The language family is sketched in Curnow and Liddicoat (1998), and the following information is distilled from their account unless otherwise indicated. Cayapa has 3,000 speakers, roughly 20% of them being bilingual, and Tsafiki 1,000 speakers (2,000 according to Dickinson 2002: 20) with about half of them being bilingual in Spanish. Grammatically, all Barbacoan languages, including Cayapa and Tsafiki, have a similar profile: they have SOV word order, are predominantly suffixing, and have alignment systems on a nominative-accusative basis. In fact, Cayapa and Tsafiki are very closely related even within Barbacoan, jointly forming the South Barbacoan subgroup of the family, and yet, they differ in the size of their consonant inventory. Cayapa has an inventory of twenty-four distinctive consonants (*p*, *t*, *tʲ*, *k*, *b*, *d*, *dʲ*, *g*, *ts*, *tʃ*, *f*, s, *ʃ*, *h*/*x*, *s*, *m*, *n*, *ŋ*, *ɲ*, *r*, *l*, *ʎ*, *w*, *j*, and *ʔ*, with the phoneme /g/ being marginal and proba-

bly introduced into the language with Spanish loans). In contrast, Tsafiki only has fifteen consonants (*p*, *t*, *k*, *b*, *d*, *ts*, *Φ*, *s*, *h*/*x*, *m*, *n*, *r*, *l*, *w*, *j*, and perhaps *ʔ*, cf. Moore 1972). In both cases it must be noted that the languages are phonologically quite underanalyzed and the systems in their synchronic state are therefore somewhat insecure. The reconstructed Proto-Barbacoan phoneme inventory, from Curnow and Liddicoat (1998: 401, table 9), is in table 18.

	Consonants				Vowels	
p	t		k	i	ɨ	u
	ts					o
Φ	s	ʃ	h		a	
m	n					
	l					
	r					
w	j					

table 18: Proto-Barbacoan phoneme inventory (Curnow and Liddicoat 1998: 401)

Thus, rather than Tsafiki having shrunk its consonant inventory when compared with Proto-Barbacoan, it is rather the case that the Cayapa inventory expanded by phonemicizing erstwhile allophonic differences which are still observable synchronically in Tsafiki. Moore (1962) reconstruct both palatal and alveolar series for Proto-South Barbacoan, i.e. the common ancestor of Cayapa and Tsafiki, which Curnow and Liddicoat (1998: 400) show to be unnecessary. Allophonic variation was phonemicized when Cayapa collapsed *o and *u, leading to the emergence of a new series of palatal consonants *ʃ*, *tʃ*, *ɲ*, *ʎ*, *tʲ*, *dʲ* next to the alveolar series which all Barbacoan languages feature (Curnow and Liddicoat 1998 leave the development of voiced stops, *s* in Tsafiki, and both *r* and *s* as well as that of the palatal stops in Cayapa unaccounted for, but this does not alter the synchronic observation that these contrast exist in the present-day languages, no matter how they arose). Moreover, *ʃ becomes *s* in Cayapa and is also lost in Tsafiki.

Not discussed explicitly by Curnow and Liddicoat (1998) is syllable structure. In Tsafiki, there are only CV syllables (Moore 1972: 76), and vowel sequences are separated by epenthetical glottal stop, i.e. there are no diphthongs (Dickinson 2002: 34). In Cayapa, CVC syllables are allowed, but the final consonant can only be a nasal, liquid, or glottal stop (Moore 1962: 273 also reconstructs this state of affairs for the common ancestor of Cayapa and Tsafiki). Reconstructions by Curnow and Liddicoat (1998: 392, table 1) suggest that Proto-Barbacoan allowed for CVC syllables, with little apparent restrictions on which consonant can be present in coda position. The coda restrictions in Cayapa are explained by the loss of word-final stops in both Cayapa and Tsafiki.

Summing up, syllable and root structure were simplified somewhat in Cayapa, while Tsafiki has shifted entirely from moderately complex CVC syllables to permitting simple CV syllables maximally. Concomitantly, the Cayapa consonant inventory expanded, and that of Tsafiki underwent some changes when compared to the Proto-Barbacoan state, but remained largely constant in terms of sheer size. Both facts converge in the same predictions about the lexicon in line with the typological evidence: expansion of the

Cayapa consonant inventory may have caused the number of its analyzable terms to shrink, while Tsafiki should have expanded the degree of analyzable terms in its lexicon due to the shift from CVC to CV structure of the syllable/root. The monosyllabic structure of many lexical roots inherited from the Proto-Language remained intact in both languages (albeit they are of different complexity, see above).

Looking in more detail at the individual analyzable terms in Cayapa and Tsafiki, there are a number of meanings expressed by analyzable terms in both languages. Some of these have the same internal semantic structure, so that, other things being equal, they should probably be taken to be inherited at least from the immediate common ancestor. These include 'guts' (Cayapa *pe-shilli* 'exrement-line,' Tsafiki *pe-silí* 'excrement liana/rope') and 'nostril' (Cayapa *quij'juru ~ quijuu ~ quij'jura* /quijcapa-juru/ 'nose-hole,' Tsafiki *quinfu foró* 'nose hole'). Further terms that are quite similar in their internal structure are those for 'ashes' (Cayapa *ñiipe* /ñi-pe/ 'fire-excrement,' Tsafiki *nin fu* 'fire feather/body.hair') and 'brain' (Cayapa *mishpe* /mishu-pe/ 'head-excrement,' Tsafiki *fu-pe* 'hair-excrement'). In the case of other meanings, both languages have analyzable terms, but with different structure.

There are also meanings which are expressed in Cayapa by an analyzable term, but not in Tsafiki. For instance, Cayapa has *ya-tape* 'house-grass' for 'nest' and Tsafiki the unanalyzable *ta'sén ~ ta'sín*. However, it is much more frequently the case that it is Tsafiki which features an analyzable term, whereas the Cayapa counterpart is either totally unanalyzable or semianalyzable. For instance, Cayapa has *ujtupe* 'dust,' Tsafiki *to poyó* 'earth/soil smoke/cloud/steam,' Cayapa has *pusu* 'lake' (< Span. *pozo*?), Tsafiki *hua pipilú* containing *hua* 'big' and *pi* 'water, liquid, river,' Cayapa has *ingbi* 'saliva,' Tsafiki *pi'pí*, presumably reducplicated from *pi* 'water, liquid, river,' etc.

Given that the relevant languages started out as being the same language (or dialect continuum) with the same or highly similar phonological and lexical structure, this is evidence that there is some structural pressure working in diachrony that causes the lexicon to adapt to subsequent phonological developments at some point of time after breakup of the proto-languages, and this appears to be the case not just in Barbacoan, but given similar results in other families, also elsewhere. What is the nature of this pressure? By asking this question, the discussion enters into the last phase of the progress towards an explanation in terms of Bybee (1988): first, *empirical generalizations* were made concerning interdependencies between four morphophonological factors, then, not the least by the computation of the CPCIs, a *principle* was formulated that summarizes several empirical generalizations, and now, the principle needs to be accounted for and an *explanation* for its operation must be sought for.

5.4.2.10. Towards a functional explanation

5.4.2.10.1 *Narrow explanation in terms of homonymy avoidance*. Linguistic universals, and for that matter, presumably also universal tendencies and correlations, are ultimately diachronically motivated and the outcome of some sort of structural or cognitive pressure pushing languages to behave in certain ways, but not in others (Greenberg 1978, 1995, Bybee 1988, Payne 1990, Haspelmath 1999, Bickel 2007, 2008). As Bybee (1988: 351) says, "synchronic states must be understood in terms of the set of factors that create them."

What the case study of Polynesian, the discussion of the situation in Proto-Indo-European, and especially the case study of Mandarin Chinese have shown is that small phoneme inventories or inventories in the process of shrinking may cause problems due to the reduced expressive possibilities, and, in drastic cases, a high number of homonyms in the lexicon when viewed in synchrony, or, when conceived of from a diachronic point of view, the creation of homonyms from erstwhile distinctive lexical items.

In fact, there is a principle said to work against this, namely homonymy avoidance: "[a]ny change in which homophony (words with different meaning sounding the same) is avoided or eliminated" (Campbell and Mixco 2007: 20). Homonymy avoidance is invoked in both synchronic phonological studies to motivate the presence of certain phonological rules, as well as in diachrony to explain aberrant phonological or lexical change, the general assumption being that ambiguity of this kind causes disturbance in the one-to-one match of form-meaning relations impeding successful communication, and that linguistic systems are designed in ways to avoid such disturbances (Plank 1981: 165, who also notes that a generalized theory of ambiguity with predictive power is lacking). The following section provides an overview of research on this, and discusses whether or not the principle of homonymy avoidance is a viable and convincing (diachronic) functional explanation for the observed correlations.

Synchronic phonological studies recurring to homonymy avoidance include the following: Awóbùlúyì (1992) demonstrates that some dialects of Yoruba, including the standard variety, have innovated a rule by which monosyllabic low-toned verbs are required to show mid tone before polysyllabic object-NPs when they are also specified for number and person. Hence, in dialects not having this rule, *mo fò díẹ̀* 'I jumped a little' and *mo fò díẹ̀* 'I skipped some' are homophonous, whereas the standard variety and relevant dialects have *mo fò díẹ̀* 'I jumped a little' versus *mo fo díẹ̀* 'I skipped some.' In Comaltepec Chinantec, a languages in which most words are monosyllabic and in which tone therefore has a high functional load in the lexicon, tone sandhi is rampant, and yet sandhi processes are almost always allophonic and do not neutralize contrastive values required to maintain distinctiveness of lexical items (Silverman 1997). Similarly, in Korean, where neutralization of contrasts is pervasive, these create a very small amount of homophony, and other plausible and phonologically natural neutralizations that would have such an effect are not part of the phonological system (Silverman 2010). In both cases, Silverman explicitly argues that the phonology is sensitive to contrast maintenance. Accounts in terms of avoidance of homophony frequently pertain to grammatical paradigms rather than the lexicon per se. In the Trigrad dialect of Bulgarian, vowel lowering in unstressed syllables is blocked if grammatical endings are present which would produce homophony (Crosswhite 1999). According to Lyovin (1977), in Classical Tibetan, gaps in verb paradigms occur when the future form would be homophonous with the present form, and that such clashes are avoided by the use of periphrastic constructions. In Carrier, diachronic vowel syncope was inhibited in a valency prefix which would have caused it to become homophonous with another one (Gessner and Hansson 2004), and in Banoni (Austronesian), erstwhile distinctiveness of vowel length was gradually lost, except to maintain distinctiveness of bare nouns and their possessed (1st person) counterparts (Blevins and Wedel 2009: 152-154).

The amount of diachronic studies in which homonymy avoidance plays a role is even more numerous. In spite of the Neogrammarian claim by Osthoff and Brugman (1878: 107) that "[m]assenhaft Beispiele beweisen ... dass die Sprache niemals aus Scheu vor Formenzusammenfall oder um Formendifferenzierung zu erhalten Lautgesetze in ihrer Wirkung inhibiert" / "copious examples prove ... that language never inhibits sound laws in their operation for fear of collapse of form or to maintain differentiation of forms," there is a wealth of literature attempting to demonstrate that just this is the case, leading Campbell (1996: 77) to state that avoidance of homonymy as a functional principle in diachrony "is an undeniable empirical reality." The case for such a principle was first made (or at least first popularized) in an oft-quoted study by Gilliéron and Rocques (1912), see Williams (1944: 23-44) for discussion of still earlier precursors. These authors famously observed that, while reflexes of Latin *gallus* 'rooster' are found throughout Southern France, they are notably unattested in Gascony. Here, 'rooster' is denoted by terms that originally meant 'pheasant' or 'vicar' and the inherited word is lost. Now, the area where this lexical replacement has taken place coincides very well with a sound change merging word-final [l] with [t]. Due to this change, Latin *gallus* 'rooster' would not be reflected as *gal*, as in most other areas, but as **gat*, which also happens to be the regular reflex of Latin *cattus* 'cat,' and their argumentation is that this replacement is motivated by the avoidance of homonymy between 'rooster' and 'cat,' two meanings expressed both by nouns and likely to co-occur in the same (rural) setting, thus endangering the successful transmission of information. Similarly, Öhmann (1934: 40) attributes replacement of *fliegen* 'to fly' by *fahren* 'to ride, go' in some varieties of German to clash with *fliehen* 'to flee.' Williams (1944) discusses, next to a wealth of other cases, the fate of English *ear* 'ear' vs. *near* ~ *ear* 'kidney.' Simplifying Williams's more complex discussion, in Northern England and Scotland *lug*, a word of unclear provenience, came to be used for 'ear,' while Standard English *ear* swamped out *nere* ~ *near*, not in the least due to additional confusion when *ear* is preceded by the indefinite article. Discussion of Dutch examples is in Kieft (1938); such early discussions are heavily inspired by Gillierón and Rocques (1912), and indeed their account has spawned much literature that attempts to unravel similar cases in other languages. Dworkin (1993a, b) shows that in Old Spanish, one of two competing same- or similar-sounding lexical items in the same syntactic category and with similar or opposed meanings were lost, and Malkiel (1952) discusses cases on the basis of data from Spanish and other Romance languages. In Proto-Aztec, reflexes of Proto-Uto-Aztecan *$t\dot{\imath}^{n}$ 'stone' and *$t\check{a}$ 'fire' would have been expected to fall together in **te* due to merger of the vowels *$\dot{\imath}$ and *ă. However, the actually attested reflexes are *tle-* 'fire' and *te-* 'stone' in some dialects and *ti-* 'fire' and *te-* 'stone' in others (namely those lacking *tl*), and this case of irregular sound change is explained by Campbell (1975), from who the discussion is summarized, by appealing to the principle of homonymy avoidance. Campbell and Ringen (1981) and Campbell (1988) provide an overview of further cases from the literature where homonymy avoidance is claimed to cause lexical loss or replacement, such as loss of Middle English *quean* 'low woman' after merger of of middle english [ɛ:] and [e:] due to conflict with *queen* except for dialects of the Southwest where the vowels did not merge (taken from Menner 1936: 232-233), replacement of *fliege* 'fly' by *mücke* 'gnat' in dialects of German because of homophony with *flöhe* 'fleas' (taken from Bach 1969: 168). Exceptions in sound

changes in grammatical paradigms are also at times attributed to homonymy avoidance, cases in point being the non-systematic retention of intervocalic *s in Classical Greek due to its function as a marker of the aorist (Bloomfield 1933/1984: 362-363) and that of word-final *n as a marker of the 1st person singular in northern Estonian (Raun and Saareste 1965: 62); for summarizing discussion of both see also Blevins and Wedel (2009) and Campbell (1975).[32]

Malkiel (1979: 2-3; 7) lists four possible outcomes of homonymic clashes (see also Williams 1944 for a similar typology): (i) both lexical items may simply continue to coexist, (ii) one ousts the other (as argued for by Gilliéron and Rocques 1912 and in subsequent studies), (iii) if a semantic gap can be perceived, they may merge (traditionally known as contamination, discussed also in Malkiel 1952), and finally (iv) they may differentiate in form (and meaning). Akin to the last point is the cause of irregular sound change or the blocking of regular sound change highlighted by Campbell (1975, 1996, 1998) and refined by Blevins and Wedel (2009). Furthermore, there is a fifth possible strategy, alluded to by Rédei (1970: 11): therapeutic borrowing, which involves borrowing of a word for one of the referents expressed by homonyms from a contact language. Haspelmath (2009: 50) also mentions that it has been suggested that the replacement of English *bread* 'roast meat' (< Old English *bræde*) by a loanword from French (namely *roast*, which, incidentally, seems to be ultimately of Germanic origin itself, as evidenced by cognates such as Middle Low German *rosten, rosteren* 'to roast on grate,' Kluge 2002) is motivated by homonymy with *bread* 'morsel, bread' (< Old English *bread*), though he remains agnostic as to whether this is really the functional motivation, referring to Weinreich (1953: 58) who uttered a similar opinion. However, there is a sixth apparent possible outcome, pointed to already by Öhmann (1934), and this is more relevant in the present context: creation of a disambiguating compound. Öhmann discusses the case of Middle High German *mûl* 'snout' (an inherited word) and *mûl* 'mule' (< Lat. *mulus*), which latter survived only in compounds like *mûltier* 'mule-animal,' *mûl-esel* 'mule-donkey' and *mûl-ros* 'mule-horse.' Furthermore, he points out that erstwhile *gift* 'gift' only survives in the complex term *mitgift* 'dower' while the simplex has been ousted, presumably due to conflict with *gift* 'poison.' Similarly, Williams (1944: 11-12) says: "[a] word threatened in its existence by some of the vicissitudes of language development, as, for example, homonymic conflict, may be strengthened, made unambiguous by a modifying phrase or term that is in time considered almost an integral part of the word." Note, however, that in the complex terms discussed by Öhmann, the original monomorphemic homonym is not ousted from the language. Coates (1968) presents a further Germanic case study highly relevant for the present context, inspired by discussion in Kieft (1938). There were three segmentally similar but distinct lexical items in Proto-Germanic: **þĭhstila* 'thistle,' **þinhslā* 'pole, beam, tongue' and **þehsalōn* 'adze.' These remained distinct in older stages of Germanic languages where reflexes are attested (Coates 1968: 470, table 2), but later, putative mutual influence and attrition of the near-homonyms lead to unexpected phonological changes in some daughter languages as well as the irregular collapse of two of the forms for instance in Frisian, where both **þĭhstila*

[32] Croft (2000: 66-68) discusses homonymy avoidance as a possible factor in the evolution in grammatical paradigms, but denies strong effect claiming that tolerance of homonymy is relatively high.

and **þinslā* are reflected as *tiksel*. Summarizing the general outcome of the near-homonymy of the words for the three referents from Coates (1968), there are four major strategies, all but one corresponding roughly to the ones mentioned in the literature. In four Germanic languages, there are (optional) compounds based on the inherited word of the redundant type. For instance, Dutch has *disselboom*, with *dissel* the reflex of the inherited **þinslā* 'pole, beam, tongue' and *boom* meaning 'tree.' Five languages have resorted to borrowing of a cognate term that is however phonologically distinct to avoid homonymic clash, for instance Swedish has borrowed *dexel* from German. In six languages, semantic change has taken place, either by replacement of inherited terms with more general meanings or by metaphorical extensions of other words. However, importantly, in some languages, terms for one of the meanings have been given up entirely, and replaced by compounds that do not involve one of the inherited words as constituents. For instance, Dutch has *dwarsbijl* 'cross axe,' Yiddish and Faroese have *bonders hak* and *bøkjaraøks* 'cooper's axe' for 'adze,' and Icelandic has *vagnstöng* 'wagon-pole' for the second of the conflicting meanings; Norwegian and Danish have analogous compounds.

Coates (1968) is important in the present context for another reason: he argues that not only perfect homonymy may be a factor, but that near-homonymy is sufficient in some cases to trigger linguistic changes such as lexical replacement, and if this is true on a larger scale, then lexical replacement due to similar forms becomes a more attractive functional explanation to account for the high numbers of complex terms in languages with morphophonologically simple systems, because, while the number of true homonyms may still be limited, the number of phonologically similar lexical items can be expected to be exponentially larger.

Still, there are serious problems in the cogency of applying the complex of data revolving around homonymy or near-homonymy directly as an explanation for the observed correlations. The first question is how disastrous the effect of a particular sound change (in particular phonemic mergers) can be for lexical distinctiveness. This would require detailed investigation of the functional load of a particular phonemic contrast within the lexicon, and to show where this distinctiveness is encroached on by the loss of the contrast. It is intuitively clear that one particular phonemic merger will not affect the lexicon as a whole, but only a well-defined subset. However, as the diachronic studies cited above suggested, already one sound change can lead to changes in the lexicon if it affects a sensitive point therein, namely lexical items that are useful for successful communication (which is after all the job of language) to be kept distinct. Then, clearly, as Lyovin (1977: 121) says in general, the more homophony is produced by sound changes, and the more dramatic they are, the more likely it is by sheer probability that they cause erstwhile distinct lexical items somewhere in the lexicon to collapse even in the same syntactic and semantic class.

One of the two more severe reasons for some skepticism, however, is that, while there are studies that show empirically that semantically redundant complex terms are introduced for the purpose of disambiguation, there is, with the exception of Coates's study, little evidence in the literature that inherited lexical items are given up because of homonymy and replaced by entirely new complex lexemes that do not contain the inherited homonym as one constituent. A further brief but notable comment is that by Shi

(2002: 76) to the effect that in the process of disyllabification of the Mandarin lexicon discussed in § 5.4.2.3.2. "[m]onosyllabic words are replaced by newly created disyllabic words, in other words, earlier monosyllabic words are abandoned," such as *yue* 'concise' by *jian-yao*, and *wu* 'understand' by *li-jie*. Further empirical demonstration of the pervasiness of such developments would be a prerequisite to make a cogent case for avoidance of homonymy or near-homonymy to account for the observed patterns. To be sure, absence of discussion in the literature does not entail absence of the phenomenon and a case made *ex nihilo* cannot be very strong, but if this were a very frequent process, one would assume that it would have been commented on by historical linguists.

What is more, there are also irregular changes that run counter to the putative principle of homonymy avoidance. Dixon (2004: 71) mentions irregular nonce changes from Proto-Arawá to Jarawara, a language which according to him, and in principle in accord with the hypothesis of phonological constraints on the shape of the lexicon, already has a high number of homonyms due to the phonological structure (11 consonant phonemes, four vowels, (C)V syllable structure, yielding 47 possible distinct syllables and thus, given the preference for bimoraic lexical roots, 2,209 possible disyllabic items as calculated by Dixon). For instance, Proto-Arawá had the distinct nouns **ino-ni/ino-ne* 'tooth,' **ini-ni/ini-ne* 'branch,' and **oni-ni/oni-ne* 'name' (suffixes distinguish masculine and feminine forms respectively). By regular change, **ino-ni/ino-ne* 'tooth' became *ini/ino* in Jarawara. However, the reflex of 'branch' is not the expected **ini-ni/ini-ne*, but *ini/ino* as well. In other words, the feminine forms of 'tooth' and 'branch' collapsed phonologically, and according to Dixon, the masculine form of 'tooth' was analogized causing lexical distinctiveness to cease entirely. Furthermore, by the normal diachronic changes **oni-ni/oni-ne* 'name' first became *oni/oni*, but has then undergone irregular metathesis of the masculine form, giving Jarawara *ino*, yielding homonymy of the masculine form with both 'tooth' and 'branch,' and subsequent extension led the feminine form *ini* to extended to cover 'name' (Coates 1968: 473 also observes that "in some cases a minimal distinction is not felt to be worth preserving, that it is regarded as no better than no distinction at all"). In fact, Dixon (1999: 297) even states that "[o]ne characteristic of Arawá languages is a profusion of lexical homonymy, in which speakers appear to delight," and if this is indeed the case, this delight is of course detrimental to the hypothesis that homonymy avoidance is a cross-linguistic valid motivating factor in language change. It cannot be entirely excluded that the wealth of studies on homonymy avoidance as motivating linguistic change are science-historically a result of the seminal study by Gilliéron and Rocques (1912) that sensitivized linguists to the issue and to look for similar cases in other languages (note also that several later authors, e.g. King 1967, called into question the pervasiveness of homonymy avoidance in diachronic change and the existence of therapeutic language change as put forward by Prague circle linguists, while often acknowledging that some changes may be due to homonymy avoidance or more generally are therapeutic measures).

The other great difficulty is that, as Hanks (2000: 206) has it, and as several of the above cited studies (e.g. Williams 1994, Dworkin 1993a,b) emphasize, sheer identity in form between two lexical items does not necessarily constitute a problem, since if they belong to different parts of speech and are semantically remote from each other, they are unlikely to constitute a danger of confusion in actual discourse, so that normal adult

speakers are unlikely to propel therapeutic measures unless perhaps the above requirements are fulfilled, which should be relatively infrequently the case.

A related issue is the personification of "language" as a deliberately acting agent inherent in some accounts (see also King 1967: 850 for critique). It is important not to forget that this is only a metaphor, and that actual speakers, not languages themselves, are the instigators of language change. In this context, a question one must ask is that if homonymy or near-homonymy is avoided cross-linguistically, by whom is it avoided? If it does not seem attractive that native adult speakers should be responsible for linguistic change caused by homonymy or near-homonymy for the above mentioned reasons, it is worthwhile to look at other groups of speakers. Trudgill (2002, 2004), for instance, argues that a considerable amount of homonymy in the lexicon is indeed tolerable for the native speaker, but is unequally more problematic for the language learner (this is part of the argument developed by Trudgill to account for the shrinking of phonemic contrast in Polynesian), since "[t]he less there is to remember, the easier language acquisition is" (2002: 714), which is also taken by him to be the reason of reduced vocabulary size in Pidgins (cf. §§ 5.4.2.12.1. and 5.4.2.12.7).[33] However, when testing for the presence or absence of second language learners by a Mixed Model design (data are in Appendix C), no statistical effect on the difference of analyzable lexical items can be observed on a global scale. Perhaps with more fine-grained systematic data which distinguishes more subtypes and detailed scenarios paying more attention to the sociolinguistics of the language contact situation etc. significant patterns would emerge, but for the time being, there is no evidence for the sheer presence or absence of second language learners on the degree of analyzability.

In contrast, there is evidence in the recent literature on language acquisition that children in learning their L1 have surprising difficulties with homonyms. To be sure, understanding the concept of homonymy requires a lot of cognitive infrastructure. Most importantly, the child has to be able to understand that a referent and the word denoting it are not the same thing and are associated to each other only by convention, and that the conventions are sometimes such that one word may have two (or more) different referents. The relevant infrastructure is developed by age four (Doherty 2004) and yet children have surprising difficulties in experimental settings with homonyms (Mazzocco 1997, Doherty 2004, see further references to earlier literature therein). The difficulties with homonyms may last as long as until the childrens' 10th birthday, that is, until first language acquisition is nearly complete. Striking examples illustrating this are provided in Campbell and Bowe Macdonald (1983). For instance a girl at age 4;3 is shown a number of pine cones and is asked "What are these things?" by an interviewer. The child volunteers the correct answer "Cones." However, when further asked "Where d'you get cones?," the girl answers "At the shop," and when asked, specifies "At Daddy's shop." This answer is surprising, but becomes at least understandable when one knows that her father is the owner of an ice cream shop. Thus, even though the girl clearly knew of the two different referents of *cones*, as evidenced by her volunteering the answer as to the name of the pine

[33] Furthermore, Trudgill (2002) argues that adult bilingualism and learning leads to phonemic simplification and child bilingualism to phonemic borrowing.

cones, she still somehow failed to keep pine cones and ice cream cones apart.[34] Now, children are also extremely creative at making up new words to fill lexical gaps when they have not yet learned the name of an object (e.g. Clark 1981, 1982, 2000, Clark and Hecht 1982). Thus, classical examples of blocking of the application of word-formation rules by an already existent word, such as the lack of an agent noun **better* 'someone who bets' by *better* and **letter* 'landlord' by *letter* (taken from Jespersen 1942: 231) may be suspended for children, in the spirit of Paul (1880/1966: 251): "Die Individuen, welche das Neue zu dem Alten gleichbedeutenden hinzuschaffen, nehmen in dem Augenblicke, wo sie dieses tun, auf das letztere keine Rücksicht, indem es ihnen entweder unbekannt ist, oder wenigstens in dem betreffenden Augenblicke nicht ins Bewusstsein tritt" / "the individuals adding the new to the synonymous old are not considerate of the latter in the moment they do so, it either being unknown to them or at least not entering their conscious mind in the moment in question." For instance, Panagl (1976) reports a child acquiring German deriving the verb *pfeilen* from *Pfeil* 'arrow,' and from *pfeilen*, in turn, the instrument noun *Pfeiler* for 'bow.' Now, if children have problems with homonymy, then it would be logical to hypothesize that it is them who suppress one of the meanings of a homonym by replacing it with a complex novel term which they coin frequently and productively in language acquisition anyway, and if these are taken over by the parents and become institutionalized, then children could be thought to be the propagators of novel descriptive terms, and in the end be identifiable as the agents propelling the correlations in the lexicon. However, to be sure, this scenario is highly speculative, and operates with the unproven assumptions that (i) children in actual life rather than in an artificial experimental setting really have problems with acquiring homonymy as well - here the same observation as for adults may well hold, namely that homonymy may not be a problem for children as well as long as homonyms do not co-occur in the same context and that the words are thus acquired in different conceptual frames with the child possibly not even realizing homonymy, and (ii) that parents propagate children's innovation through the speech community and thus procure conventionalization of the putative innovations.

Since there is no cogent evidence for child or adult language learners as agents in linguistic change with respect to the topic discussed here, it is appropriate to return to adult native speakers for a moment. It is not always the case that speakers can be sure that context will resolve ambiguities in their messages. Charles-Luce (1993, 1997) demonstrates that phonological processes in speech production are sensitive to semantics and pragmatics of the context, in particular that phonemic contrasts in lexical items are preserved more faithfully in semantic and pragmatic contexts where the speaker cannot expect the listener to expect the word to occur in discourse. Shields and Balota (1991) report that the duration of a target word in a sentence was shortest when the target word had already occurred in the same sentence before, somewhat longer when a semantically related word had occurred in the sentence, and longest when the target word was not related semanti-

[34] Note that in a linguist's analysis, the case of *cone* might be treated as a case of polysemy created by metaphorical extensions from '(pine) cones' to '(ice cream) cones' due to similarity in shape. However, for ordinary language users it may be the case that they would not perceive any semantic link between the two referents, in other words, that for them it may be a case of plain homonymy rather than polysemy.

cally to another one having occurred earlier in the sentence. Likewise, Fowler (1988) shows that words repeated in the same stretch of discourse are, compared to unprimed occurrences, shortened in their pronunciation duration by the participants of her experiments, but not when the words are read from a list, and, very importantly for the present context, neither when the words are preceded in discourse by homophonous items, in which case their pronunciation duration is in fact somewhat longer (Fowler 1988: 313)! What this shows is that speakers are aware of potential ambiguities in communicative contexts and actively (though perhaps subconsciously) take countermeasures to make sure to be properly understood, and it may be precisely this fact that is in the end responsible for diachronic effects of homonymy or near-homonymy: irregular sound change due to overly careful, exaggerated pronunciation of relevant lexical items, or their replacement in ambiguous contexts by semantic proxies potentially leading to semantic shift as conventionalization sets in, or their replacement by a circumlocution, which is the most relevant aspect for present purposes. This would be a step to solve the first problem noted above, namely that homonymy is only pernicious if the relevant items belong to the same part of speech and the same semantic domain, as well as the problematic likening of language to a deliberately acting agent.

This explanation is compatible with the old proposal that speakers are caught in between to opposite drives: on the one hand to avoid unnecessary articulatory effort in order to not waste energy, but at the same time have to make sure to be properly understood (Gabelentz 1901, Martinet 1952, Haspelmath 1999, among others). And if these findings are replicable cross-linguistically, then there is a way to escape Haspelmath's (1999) teleological fallacy to take "functional statements as sufficient explanations," by tying the functional statement up with speaker behavior: languages do not have many analyzable terms in order to counter reduced phonological resources, but because speakers introduce them to ensure successfulness of communicative events. However, the main obstacle for a more detailed fleshing out of the precise workings of the principle is at this point of time that, as noted by Geeraerts (2002b: 37) "actual research into homonymy at the level of *parole* is scarce." For this reason, an in-depth discussion which zooms in from the typological bird's eye view to exemplary studies of the actual processes that might operate in discourse to bring about homonymy avoidance is unfortunately scarcely possible.

5.4.2.10.2. *Broad interpretation as a functional continuum.* Apart from homonymy avoidance per se as the functional drive, it is perhaps worthwhile to conceive of actual homonymy, that is, total formal identity, as only the tip of the iceberg of a larger, but less specific pressure exerted by phonological and morphological factors. The evidence presented here suggests that limited phonological resources cause languages to exploit word-formation devices to build their vocabulary to a greater extent. This, in particular when keeping in mind the obtained correlation between the size of the consonant inventory and root structure, is entirely in line with Nettle's (1995) finding of a correlation between size of phoneme inventory in general and mean word length on the basis of a sample of ten languages: the more phonemes the shorter the words, the less phonemes the longer the words. But Nettle does not take into account whether the words have internal morphological structure, so analyzability in the sample lexical items is likely to contribute to Nettle's

findings to some degree, next to the correlation between canonical structure of the lexical root and consonant inventory size mentioned above. In fact, Nettle (1998: 244) argues that "lexical expansion" as a mechanism of adaptation is responsible for longer words in languages with smaller phoneme inventories, which draws near or is even identical to the coinage of morphologically complex words, which are of course, next to being morphologically complex, also longer.

Maddieson (1984: 8) devotes some discussion to possible effects of simplicity in phonological structure for contrastive possibilities, taking the position that there is little evidence of such effects which would include either "unacceptably high incidence of homophony or unmanageably long morphemes." By inspecting dictionaries of languages with very small consonant inventories, amongst them Rotokas and Hawaiian, he concludes that no such consequences on the morphemic level are discernible. An earlier version of the Hawaiian dictionary that is used also for the present purposes, according to Maddieson (1984: 8), states in the preface that the average number of phonemes per morpheme is just 3.5, which Maddieson finds "clearly not unacceptably long." However, note that Maddieson's discussion is concerned with the level of the morpheme, not that of the lexical item, and the properties of the lexicon suggest that here, to some extent morphologically complex items are used for purposes of disambiguation.[35] This is also supported by some amount of semantically redundant complex lexical items in Hawaiian. For instance, *ake* means 'liver' as well as 'to desire, wish, be eager, yearn.' The meaning 'liver' can be singled out by using the compound *ake-paʻa* 'liver-firm' which is "more specific than *ake*" according to lexicographers (and note that 'lungs' in Hawaiian are either called *ake-māmā* 'liver-light,' *ake-makani* 'liver-wind,' or *ake-pāhola* 'liver-spread'). For present purposes, they do not affect the outcome, since complex terms such as *ake-paʻa* are treated as being redundant and are not taken into account, and effects may be more dramatic if they were.[36] A general tendency one would expect on the basis of the correlations established here is that such formally redundant terms are more frequent in languages with low phonological complexity, essentially serving the same purpose as non-redundant complex lexical items, namely to increase lexical distinctiveness when necessary. This would require further testing.

It is also instructive to look at the ratio of potential words that can be generated by the phonological system and the ones actually instantiated. According to Krupa (1966), in Maori, which has a slightly larger phoneme inventory than Hawaiian and Samoan, the

[35] Trudgill (1996: 15) interprets Maddieson to the effect that "it is not the case that languages with small inventories necessarily have longer words, or vice versa," but note again that Maddieson is talking about the morpheme, not the word.

[36] While it is theoretically conceivable to simply add segments to words arbitrarily to enhance distinctiveness, it seems unlikely that an actual speech community, faced with a limited number of acceptable word shapes due to phonological restrictions and canonical roots shapes, agrees by convention to add a sequence of meaningless phonemes to pre-existing words just to increase their distinctiveness. To do so, a much more natural tool is available, namely that to employ the language's word-formation mechanisms to form compounds on pre-existing roots (which then yields a large amount of semantically redundant compounds, as in Mandarin Chinese), or to replace parts of the stock of inherited words by morphologically complex neologisms.

theoretical number of (C)V syllables is 55. 38 of them are attested, representing 67 morphemes which all express grammatical meaning. From these data, Krupa, inspired by similar indices in Greenberg (1960), derives a so-called index of homonymy of 1.76 by dividing the number of morphemes with distinct meanings through the number of syllable shapes. The theoretical number of bi-vocalic morphemes, which presumably bear mostly lexical meaning given the disyllabicity of Austronesian lexical morphemes (Blust 2007), is 3,025. Of these, 1,258, that is, 41.59%, are actually observed. The index of homonymy calculated by Krupa on the basis of a chance sample of 100 items is 2.27, that is, each lexical morpheme in Maori has on average more than two distinct meanings. What this shows is that it is not necessarily the case that in languages with simple phonological systems all potential word shapes are lexically exploited in spite of many lexical items being homonyms (although there are languages where the ratio of attested to possible word shapes is higher, for instance White Hmong, according to Ratliff 1992). This is on the one hand hardly surprising, since after all, speakers do not engage in mathematical calculations of the number of possible words in their languages, and certainly they do not search for phoneme combinations not yet exploited lexically to immediately do so by the mysterious process of *Urschöpfung* at the next best opportunity, given that, after all, the lexical inventory is an organic whole that is for the most part inherited, not created from scratch. But this does not entail that, rather than searching for gaps, which seems unrealistic, speakers resort to the exploitation of word-formation devices for disambiguation of existing homonyms.

Suggesting pressure on the lexicon arising from phonological simplicity is not equal to postulate any principle of grammatical, lexical, or cognitive organization that generally averses homonymous items from the lexicon, homophonous morphemes from grammatical paradigms, or the development of such items in diachrony, in particular if these are limited to relatively few isolated instances (in this following Blevins and Wedel 2009). Rather, what the evidence suggests is that if phonological possibilities are restricted, and the ratio of instantiated lexical items approaches a certain percentage of all possible lexical items generatable by the morphophonological system (not necessarily even close to 100%, as Krupa's 1966 calculation shows), then there is functional pressure on the linguistic systems to develop strategies to counter the limited expressive possibilities constrained by segmental restrictions, either by the introduction of phonemic tonal contrasts (Matisoff 1973), and/or a notable and statistically verifiable increase of morphologically complex items. As Trudgill (2004: 315-316) says with respect to (unguided) second language acquisition specifically, "[t]he problem lies in the relative lack of distinctiveness between one vocabulary item and another, due to the necessarily high level of usage of all possible syllables," but it seems that this is precisely also the tendency that is observable from a cross-linguistic point of view, irrespective of whether the language is learned by L2 speakers or not.

Alternatively, rather than searching for an explicit functional explanation which is operative, the correlation may simply best be viewed as a constant equilibrium, where languages level off at some point on the complexity on the scale, the endpoints of which are extreme simplicity in the lexicon which goes hand in hand with complexity in phonology and the structure of the root on the one hand and dominant analyzability in the lexi-

con accompanied by phonological simplicity and simplicity in root structure.[37] For instance, many languages of Australia populate the niche of the continuum in which complex lexical items are few, but words correspondingly long. Preponderance of an analytic lexicon could be viewed as an attractor, a term adopted by Blust (2007) from Kelso (1995) to account for the remarkable stable disyllabicity of Austronesian languages, and the concomitant countermeasures invoked by individual languages to restore disyllabicity when it is in danger by sound changes (see also Nettle 1995 for an account of the phonology-lexicon interface similar to a self-organizing system).[38] Likely, there is another counteracting tendency to keep memory load within limits (cf. also Fortescue's suggestions discussed in § 5.4.2.12.4) pushing in the other direction: as Lindblom (1998, 2000) argues on the basis of neurological evidence, there are likely memory constraints which favor re-use of already lexically exploited articulatory movements. And then, these are the poles on the continuum of two opposed drives between which speakers of languages are suspended, and which in the end causes languages to level off at some point of the continuum, with none of them cross-linguistically favored, but jointly defining the space of cross-linguistic variation. Thus, this account would be a slightly modern version of the old notion of poles of articulatory ease and communicative efficiency between which speakers are suspended (Gabelentz 1901, Martinet 1952, Haspelmath 1999), and which cause the languages they speak to be spotted on some place of the continuum.

A similar scenario to the one developed here is also outlined by Nettle (1999: 144), who specifically also mentions the arise of homophony as a result of loss of segments in large phonological inventories due to difficulties to distinguish adjacent segments with similar, but not identical, articulatory and acoustic properties in actual discourse:

> as a result, 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.

However, as noted above, actual evidence that new words are coined to avoid homonymy in particular, is not totally lacking, but relatively sparse.

However that may be, such scenarios are interesting also in light of the recent surge in interest in linguistic complexity, beginning with McWhorter (2001) and challeng-

[37] A phonological simplicity/complexity continuum, or rather, circle, along which languages move diachronically is outlined by Haudricourt (1968). Without making reference to Haudricourt, Nettle (1999: 142-143) adds some flesh to the abstract proposal: Simple inventories develop by way of underarticulation where the communicative context permits, and complexification "through a combination of coarticulation and word truncation." When inventories become large and distinctive segments are closer together in the articulartory space, failure to distinguish adjacent segments may occur, again reducing the size of the inventory.

[38] Another such self-organization tendency in phonology is that of feature economy leading to a tendency to maximally exploit features for distinctive purposes. Feature economy typically pertains to features that are lexically distinctive (Clements 2003: 328), and thus is a phenomenon interacting with the lexicon.

ing the so-called equi-complexity axiom for instance uttered by Hockett (1958) but said to have still earlier precursors (Kortmann and Szmrecsanyi 2009: 266), according to which all languages have overall the same degree of "complexity" (how this should be defined precisely is a matter of debate and there is no apparent consensus in the literature. Miestamo 2008 distinguishes two basic readings: the absolute one, where complexity is taken as an objective measure characterizing the linguistic system, and the relative one, where complexity is equaled to cost or difficulty for language users, both native and non-native; see also Nichols 2009a for suggestions), with "complexity" in one area of the grammar (say, morphology) balanced by simplicity in another (say, syntax). Nichols (2009a), drawing on a typological sample, reports having found neither evidence for a preferred level of linguistic complexity in her metric, nor for a functional trade-off between complexity and simplicity in different sub-domains of grammar. The correlations found here can, if one wants, be construed as evidence not for such a functional trade-off in complexity between different subsystems of grammar, but between phonology, root shape, and lexicon, at least if one is willing to equate morphological complexity in analyzable lexical items with "complexity" in one of the senses used in the recent literature, and if one agrees, which is perhaps less controversial, to calling languages with small consonant inventory systems, simpler syllable structure, and shorter lexical roots, more "simple" morphophonologically than languages with the opposite properties.

But first, before accepting either the narrow or the broad explanatory framework for the correlations with the morphophonological factors as explanatory, it is of course necessary to consider possible alternative hypotheses, and to see whether any fares better. This entails looking at both other correlations with structural factors emerging from the preliminary tests based on WALS, as well as other other possible explanations that come to mind, to see whether there are serious alternative explanations available.

5.4.2.11. Other significant correlations with WALS

In § 5.4.2.1., preliminary tests suggested interactions with a number of WALS features, of which so far only those pertaining to phonology have been discussed in greater detail. As for the other features, Mixed Models taking into account areal factors lended no support to an interaction with the order of adjective and noun (p = .086) and purpose clauses (p = .244). The remain features "survived" this additional control, and will be discussed in the following.

There is a very significant interaction when testing for correlations with the WALS features with the presence of possessive classification and the size of the classes in such systems when they are present. This feature remains significant when controlling for area in a Mixed Effects Model and the factor itself has significant power to predict the degree of analyzable items at p = .0004. There is an overlap of twenty-seven languages between the two samples on which the statistical test is performed; however, only two of these, Khoekhoe and Kolyma Yukaghir, are spoken outside the Americas. As noted by Nichols and Bickel (2005c), most often such systems are binary, in which case it is more widely known as a contrast between alienable and inalienable possession. Rather than seeing this phenomenon as being primarily driven by the semantics of the possessed element (for instance, kinship terms and body-part terms are semantic fields that are fre-

quently inalienably possessed), they conceive of it as being primarily lexically conditioned. Figure 30 shows that the more possessive classes there are, the higher the number of analyzable terms among those investigated.

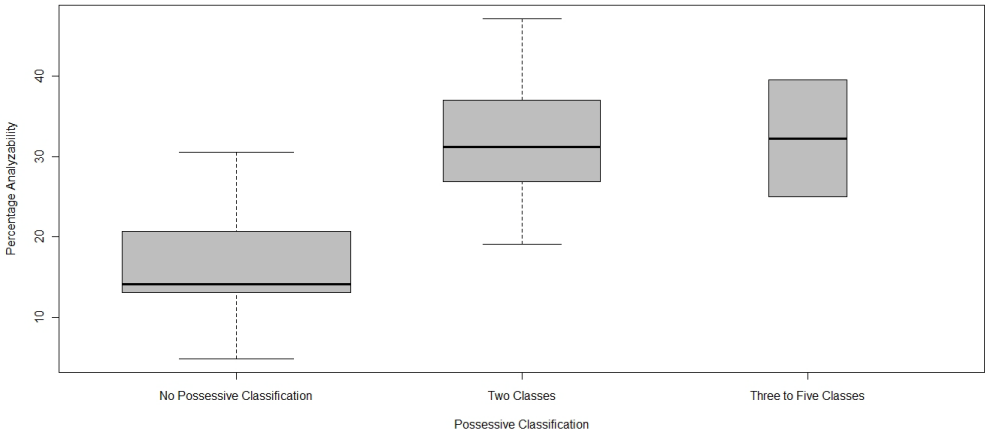


fig. 30: analyzable terms depending on possessive classification

Indeed, there is some evidence that possessive classification may be used to bring to light by morphological means different aspects of the semantics of lexical items. This point is made by Aikhenvald (2007: 38), who notes that in Tariana, an Arawak language, the same lexical item, *kare*, means 'wind' when alienably possessed and 'my breath, my heart' when inalienably possessed and prefixed with the respective marker *nu-*. However, there is little evidence from the sample data that this distinction is exploited on a larger scale to enrich the lexicon, although it is to some extent in languages of the Americas, as in Tariana. One sample language where it appears to be exploited to some degree is San Mateo del Mar Huave. Here, inalienable possession is marked by the suffix *-aran*, and the semantic domains it applies to mostly are, as is typical, body-part and kinship terms. Inalienable possession is optionally marked by the prefix *mi-* (Stairs and de Stairs 1981: 291-292, the authors do not use the terms inalienable and alienable possession, but it seems clear from their discussion that this is a typical system of possessive classification).[39] Often there is no apparent effect of the suffix for inalienable possession on the semantics of the root, as in (4.).

(4.) *mijiw-aran* 'breast/teat-INAL.POSS'

However, at times, roots bearing the suffix differ from those without it semantically, and it seems that it is used in a derivational fashion in (5.).

[39] Stairs and de Stairs (1981: 294) note that *-aran* can be added to a nominalised verb in which case the new form conveys 'clasificación,' as in *ajiüng* 'pray' – *najiüngaran* 'prayer.'

(5.) a. *mipeparan* /mi-apep-aran/ 'AL.POSS-inflate/globe-INAL.POSS' = 'bladder'
b. *omeaats-aran* 'inside-INAL.POSS' = 'heart'

A further possible example are terms for the 'testicles' in Ineseño Chumash, which consist of words for 'pit, seed' and 'stone' with the possessive prefix *is-*.

However, analysis of the data in the validation sample does not lend support to this evidence. Quite to the contrary, the estimate between no possessive classification at all and two classes is negative as opposed to positive here (-7.150) and thus not at all within that of the original sample (14.685 ± 3.134). The same is true for the features dealing with predicative adjectives.

As for the order of demonstrative and noun as a possible predictor, which is visualized in figure 31, validating the results of the original sample is difficult, because the languages in the validation sample fall in two groups only, those with demonstrative-noun order and noun-demonstrative order, with none of the rarer types involving demonstrative affixes and others with demonstrative elements on both sides of the noun mixed behavior figuring in this sample.

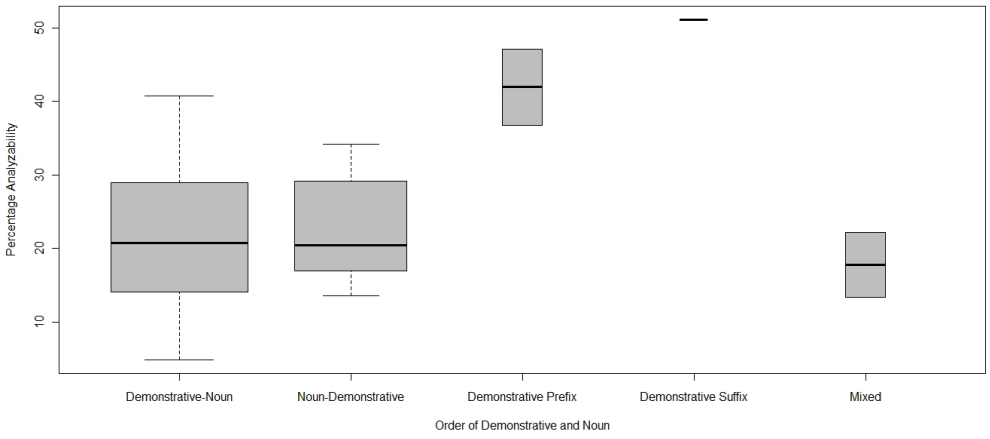


fig. 31: analyzable terms depending on the order of demonstrative and noun

As for the difference between the first mentioned major types, results are similar (1.254 ± 2.367 vs. .175), but since the drastic differences causing the original model to become significant occur with the types involving affixed (estimates are 20.367 for demonstrative prefixes and even 29.517 for demonstrative suffixes), the similarity between the results regarding the major groups is not very informative. In the original sample, languages with demonstrative prefixes are Abzakh Adyghe and Pawnee, and the one language with suffixes is Kiliwa, which all have unusually high percentages of analyzable terms. Given that there are thus only three relevant observations available, further data for languages with demonstrative affixes would be required to give a definite answer to the question whether this factor influences the behavior of languages with regard to analyzability. For now, the result is suspicious of being purely accidental.

As for semantic distinctions of evidentiality, the correlation shown in figure 32 remained significant in a Mixed Model taking into account areal factors at p = .0069, and it was possible to replicate the difference between languages with no grammatical evidentials and those with indirect evidentials, but not that between the latter and those featuring also direct evidentials (5.812 ± 3.143 vs. 6.162 and -8.015 ± 4.019 vs. 6.462).

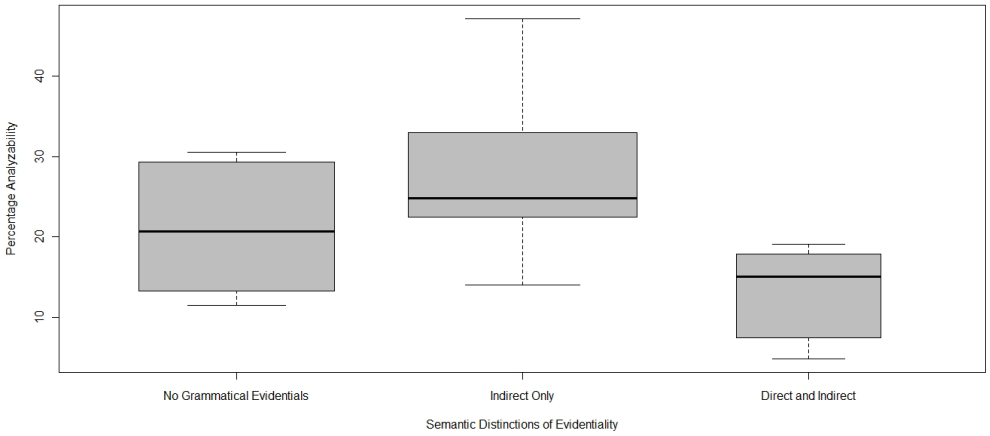


fig. 32: analyzability depending on semantic distinctions of evidentiality

Given this mixed result and the fact that it is unclear why the correlation should be there in the first place, the conclusion is that semantic distinctions of evidentiality do not seem to influence the number of analyzable lexical items to the same degree as the phonological features do, although further testing would be required to ultimately rule out a true effect.

The interation with predicative adjectives remains significant at p = .03823 when controlling for area, but the effect cannot be replicated on the dataset of the validation sample (estimates -6.539 ± 3.438 as opposed to -17.956 for nonverbal encoding and 8.836 ± 5.5 as opposed to 2.019 for mixed encoding), which suggests that the effect is not genuine.

Summing up, there is little evidence that among the structural features coded in WALS, any other than the phonological ones play a role in shaping the degree to which languages resort to analyzable lexical items.

5.4.2.12. Further tests and possible factors

5.4.2.12.1. Sociolinguistic Function: Esoteric vs. Exoteric Languages. Thurston (1989) proposes a distinction between languages with respect to their sociolinguistic function that is said to correlate with structures in grammar and in particular the lexicon. Esoteric languages, according to Thurston (1989: 556), "function primarily as codes for communication among people of the same social group. Over time, they tend to become gradually more complex. That is, they acquire a relatively high degree of allophony and allomorphy; they build large vocabularies with many near-synonyms and many opaque idioms; and they come to make relatively more numerous obligatory grammatical distinctions." Exoteric languages, in contrast, "have, as at least one of their primary sociolinguistic functions, use as a lingua

franca between peoples of different social groups. They tend to be structurally simpler than esoteric languages, because they must be easily learned by adults with different linguistic backgrounds" (Thurston 1989: 557).[40] More specifically, Thurston (1989: 567) argues that one diagnostic for exoteric languages may be "the relative lack of monomorphemic lexemes, particularly for terms that are usually considered endolexical [i.e. terms belonging to basic vocabulary]." This is, according to him, the result of coinage by second language learners when the name in the target language for a specific extralinguistic entity is lacking. Thurston states that "[w]hen more data of this sort are collected, I anticipate that a correlation will be found between the degree of esoterogeny and the number of highly specific monomorphemic lexemes."

This anticipation is open to empirical investigation using the data of the present study. In order to test Thurston's prediction against the present data, information on the sociolinguistic function of the languages in the statistics sample was gathered. In particular, attention was paid to whether the languages in the sample do have second language learners (however many) or not. These data were obtained primarily from the consulted sources for each language themselves or from Lewis (2009) and are found in Appendix C. This is a rather coarse measure, and it is acknowledged that it simplifies Thurston's more complex scenario somewhat in order to make it testable empirically. On the other hand, the coding should mirror Thurston's distinction to a reasonable degree, since exoteric speciation in his sense necessarily entails second language learners while esoteric speciation does not.

After constructing a Mixed Model design as usual, there was no appreciable difference between languages of either kind in the presence of morphologically complex terms itself (p = 0.7044). As the estimate of the model at 1.219 shows, it is even the case that languages without L2 learners have a slightly elevated number of morphologically complex terms when compared with languages that are not learned by second language learners, but even this observation is clearly not strong enough to be of significance. A visualization of the values is in figure 33.

[40] For a recent application of Thurston's dichotomy from a cross-linguistic point of view, see Lupyan and Dale (2010). Wray and Grace (2005) also heavily borrow from Thurston's work, although they speak of exoteric vs. esoteric functions of languages rather then exoteric vs. esoteric languages themselves, correctly pointing out that one and the same language may be used both for in-group communication as well as communication with outsiders.

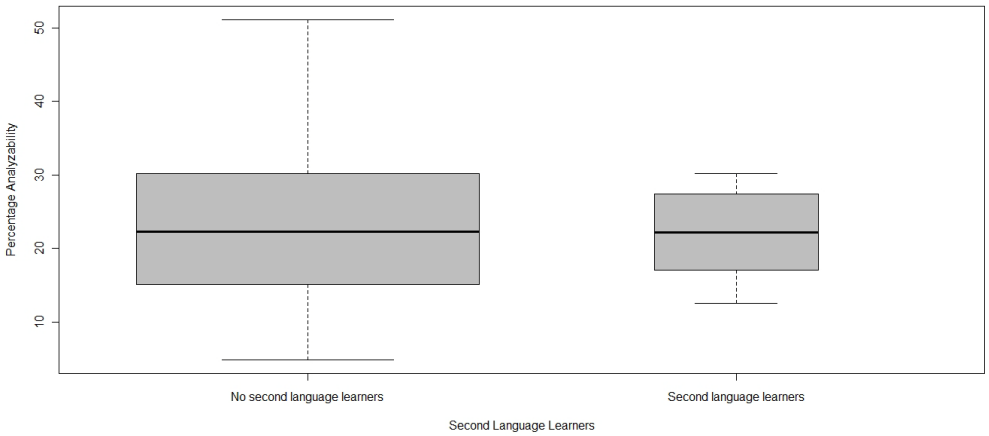


fig. 33: Morphologically complex items in languages with and without second-language learners

Nor was there a correlation with most of the other major global variables coded for each language (motivated terms in general, degree of metaphor and degree of contiguity, degree of deverbal formations).

Thus the data of the present study do not support of Thurston's expectation to find an elevated degree of complex terms in core vocabulary in exoteric languages (operationalized here as languages with second language learners). In fact, there is at least one case of a language outside the sample with sufficient published material on the matter where Thurston's predictions as to a correlation between esoteric speciation and monomorphemic lexical items do not go through. Yélî Dnye would be a textbook example of an esoteric language: it is spoken on an isolated island of the Pacific, and contact with outsiders is rare. And indeed, its grammar exhibits an enormous amount of complexity, as manifested in an elaborate apparatus of cross-referencing, often in a portmanteau fashion, with rampant morphophonological alternations, a highly suppletive verbal lexicon, and the probably most complex phonological system in the Pacific, featuring a number of cross-linguistically extremely rare sounds. This exuberant complexity causes that the language is rarely if ever successfully learned as a second language, and even women from other islands who marry a Rossel islander usually learn the language only very imperfectly (Levinson 2006a: 20-21). Still, Levinson (2006b: 230) notes that "Yélî Dnye is a language where many important, commonly employed nominal concepts are expressed with compounds." Judging from the evidence (also for body-part terms, the domain from which Thurston's original examples come) presented in Levinson (2006b), the language is not very different in this regard from Thurston's example of Anem, a exoteric language in his terms, adduced as support for the language's exoteric speciation.

On a related note, tests using the number of speakers of the languages in the statistics sample as a predictor variable for any of the major variables surveyed were carried out, with no significant results (the *p*-value for the percentage of analyzable terms is .3074

and that for the percentage of derived terms after logarithmic transformation is .2). In particular, a correlation between word length and presence of language contact and concomitant bilingualism as suggested by Trudgill (1996) on the basis of differences in word length between Standard Greek and northern dialects, which are in contact with neighboring Balkan languages on the basis of the first 50 items of the Swadesh list, could not be found on the basis of the present data on the language rather than dialect level ($W = 239.5$, $p = .6635$, Wilcoxon rank sum test).

5.4.2.12.2. (Large-Scale) Borrowing. As is obvious, heavy borrowing has the potential to have profound effects on the degree of analyzability of the lexicon, simply by the fact that in borrowed words (at least when defined strictly as the transfer of lexical material, that is, excluding calquing), possible internal morphological structure in the donor language is lost in the recipient language. Of course, derivational morphology may also be borrowed along with lexical items and subsequently nativized, as has been the case for instance in English borrowing from French, but it is probably safe to say that in most instances of borrowing, this is done at the expense of possible internal structure in the donor language. For instance, Sasse (2001: 503) appeals to the long history of mutual borrowing in languages of Europe to account for the "inexhaustible number" of simplex lexical stems found there. When it comes to large-scale borrowing, Australia also immediately comes to mind. Dixon (2001, 2002) proposes that for the most part of the continent, neighboring languages, due to extensive bilingualism teaming up with avoidance registers causing a constant need for replacement vocabulary for taboo words, on the long run end up sharing about 50 per cent of vocabulary, irrespective of genetic relatedness. This is known as the "50 per cent equilibrium model." An important study on this is Heath (1981), describing the situation in languages of Western Arnhem Land. However, this model is not universally accepted by Australianists, and the effects of word taboo are said to be overestimated by Alpher and Nash (1999). Evans (2005), while admitting that it is possible to reach figures as high as 50 per cent of shared vocabulary, adduces evidence from several Australian languages in contact, but still with undramatic levels of shared vocabulary. His conclusion is that there is significant variation in Australia in the extent of borrowing from area to area, depending also on the nature of social relations between speakers, and that the 50 per cent equilibrium model is not empirically well-substantiated on a larger basis in the Australian area.

Now, coming to the relevance of this in the present study, it is the case that two sampled Australian languages, Ngaanyatjarra and Nunggubuyu, have extraordinarily low numbers of analyzable terms, while a third, Gurindji, has somewhat more, but is shown by McConvell (2009b: 794) to feature many loanwords, around 45 per cent of all items in the World Loanword database and thus drawing close to the 50% figure Dixon's model. However, body-part terms are only moderately often borrowed. There is also one clear, instance of taboo-induced replacement, but McConvell (2009b: 797) denies strong effects of taboo on borrowing in Gurindji as proposed by Dixon.

Borrowing may also well play a role in producing the small number of analyzable terms in Ngaanyatjarra and Nunggubuyu, but another fact about these languages is that, like in many other languages of Australia, lexical roots are quite long in terms of number

of syllables. Yir Yoront, in contrast, which, unusually for an Australian language, features productive compounding, and where, interestingly and equally unusually for Australia, roots are generally short, monosyllabic or at the very least disyllabic, has the highest percentage of analyzable terms of the sampled Australian languages. This fact opens up a more parsimonious explanation, also in light of the controversiality of the status borrowing has in languages of Australia in the theoretical discussions, in that a cross-linguistically valid tendency, namely for languages with long lexical roots to have fewer analyzable terms than those with shorter roots, can be used to explain the differences in analyzability in the sampled Australian languages. This is simply the application of Ockham's razor, and it is not claimed that borrowing has no role to play, both in Australia as well as in the rest of the world. Indeed, data from the World Loanword Database can again be adduced to assess the question of interrelations between borrowing behavior and morphological complexity. Using the same subset of languages as in § 5.4.2.7.1., there indeed is a correlation at p = .02532 between the number of analyzable and borrowed terms to the effect that where there are many borrowed terms in the language, the simplicity score is higher, as seen in the plot in figure 34.

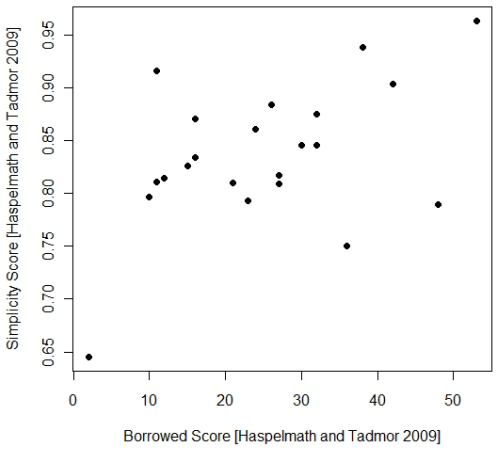


fig. 34: Correlation between borrowed and simplicity score in a subset of languages of the World Loanword Database

However, the positive correlation crucially hinges on the behavior of just one language, Mandarin Chinese, which is represented by the dot in the lower left corner in the plot in figure 34. This is relevant because for this language in particular, the differences between this study and the World Loanword Database in the assessment of analyzability is an important factor, in that complex terms of the redundant type are not motivated as defined in § 3.6.1., and thus not counted here, while they are in the World Loanword Database. If the peculiar case of Mandarin Chinese is removed from the dataset, the correlation also ceases to be significant (p = .159).

What is more, if the preliminary evidence from the case study of the Americas, which suggests that the predilection of a language for borrowing is not entirely independent of its lexical profile with respect to analyzability in native vocabulary (see § 5.4.2.7.1. for details), is valid, it enables one to treat borrowing behavior as a result of lexical organization with respect to analyzability rather than its cause.

5.4.2.12.3. Word Taboo. Taboos against naming the dead appear to be widespread around the world. Kroeber (1925/1976: 360) says that this principle is widespread in California, and, as a case in the American Northwest, Elmendorf (1951: 207) argues that in the Salishan language Tswana "the spread of derivative or compound descriptive terms through the lexicon, these terms originating as coined substitutes for tabooed words" is a likely concomitant of taboos against naming the dead, "an occasional but active custom." Since the taboo required words resembling the name of the deceased person not to be uttered, Elmendorf (1951: 207) concludes that in the course of time, the procedure would oust all words resembling a personal name. Comrie (2000) reports that in Haruai society people's names are identical to the names of everyday objects, and that at the same time a taboo against uttering the name of taboo kin is in place, which obviously leads to practical complications. For this reason, a large number of synonyms (many of them loanwords) exist in the Haruai language as a kind of backup for the event that the indigenous word should become unavailable. Of course, in a comparable situation where, unlike in the case of Haruai, neologisms are coined or related words are semantically extended rather than words borrowed from foreign languages, this would lead to a notable increase in lexical motivation. Indeed this is apparently the case in a large number of Austronesian Languages, where people's names also coincide with lexical items (Simons 1982, see also Rensch 2002: 192-195 for brief discussion of Polynesian specifically). It would be extremely interesting also to ascertain in which parts of the world people's names are at the same time ordinary words in the language with lexical meaning.[41] Further, Gamkrelidze and Ivanov (1995) attribute a large number of lexical innovations in different branches of Indo-European to taboo-induced replacement, most often in the case of names for animals which are said to have ritual significance (see also Emeneau 1948 for discussion of "hunter's taboo," forbidding to utter the name of an animal being hunted, which also seems to be very widespread cross-culturally), Tetun features a special register called *lia tasi* used while at sea fishing (van Klinken 1999: 8-10), and Barlaan (2003) discusses replacement vocabulary during the rice-harvesting season among the Isnag.

The extent of the phenomenon and its precise characteristics in different parts of the world is unfortunately not at all clear. One clear case for the presence of different subtypes of naming taboo on an entire continent is Australia (Dixon 2002), although the usually assumed far-reaching effect of this cultural practice on the lexicon of Australian languages has been challenged (Alpher and Nash 1999).

[41] This is frequently indicated for Buin in the consulted source: for instance, *kuruku* 'thunder' is a female name, while, to adduce data from a language from another area of the world, Nez Perce *símux* 'charcoal' is also indicated to be a man's name. Thus, the data of the present study indicate that the phenomenon is well attested, but are not sufficient to allow for more systematic exploration.

Testing effects of word tabooing cross-culturally is not an easy task, because no large-scale comparative anthropological treatment of patters of word taboo is presently available that would make clear just in which cultures word taboo rules are in place and where such practices are unheard of. Thus, there may be instances of erstwhile complex taboo words being conventionalized in the ordinary lexicon (and in Tswana precisely this seems to be the case to some extent). For instance, Koyraboro Senni has *taa-haa* 'sew-thing' for 'needle' to replace the monomorphemic ordinary term *sana* which must not be used at night. The question is how pervasive influence of word taboo can be on the nominal vocabulary as a whole, and whether it is strong enough to be capable of shaping lexical structures on a large scale rather than replacing single lexical items every now and then in a piecemeal fashion. This remains unclear. For the time being, it is possible to at least note that the most widespread case of taboo words reported in the literature pertain to the names of (predatory or game) animals and are thus unlikely to influence the percentage of overt marking in the meanings investigated here. Further, there are languages in the sample with both a very low degree of analyzability in the lexicon where there is no evidence for any sort of word taboo being operative. For instance, for Bora, a language with comparably many morphologically complex terms in the lexical items investigated here, Frank Seifart (p.c.) reports that any practice of word tabooing is unknown to him. Likewise, Zaira Khalilova and Madzhid Khalilov (p.c.) report no evidence for practices of word taboo in Bezhta, a language with few analyzable terms. Explicit statements in the literature for the absence of word taboo are unsurprisingly rather hard to come by, but Epps (2008: 15) mentions that Hup society, speaking a language with a relatively high degree of analyzable terms is egalitarian and liberal, with few social taboos and restrictions. This shows at the very least that presence or absence of word taboo cannot be the single underlying cause of differences in analyzability on the lexicon. Also, the outcome of taboos may be quite different: either it can lead to (massive) borrowing, as stated for Australia with its widespread use of replacement registers for certain kin relations, or it can lead to descriptive neologisms to replace the tabooed lexical item, so that it could influence the lexicon theoretically in either way with respect to the degree of analyzability.

5.4.2.12.4. Syntheticization. An explanation of increasing morphological complexity in the lexicon not directly related to phonological factors, but appealing to learning difficulties of great allomorphic variation is offered by Fortescue (1992) for polysynthetic languages[42] in general, using Eskimo-Aleut as his example.[43] Fortescue (1998: 49) summarizes that

[42] Fortescue is aware of the difficulties in defining polysynthesis (cf. also § 4.5.1.2.2.), and says that polysynthetic languages have traits that in sum allow "the expression within complex word forms of numerous elements that in more analytic languages correspond to independent lexical items, verbs thus often corresponding to whole sentences in the latter"(Fortescue 1992: 242fn1).

[43] Proto-Eskimo is reconstructed by Fortescue et al. (1994: xi) as having fifteen native consonant phonemes, with many additional non-native ones due to borrowing. It allowed maximally for CVC syllables. Canonical stems have (C)Vt(ə)- and (C)V(C)CV(C)- shape (and there are possibly also corresponding trisyllables) with some phonotactic restrictions as to what consonant may appear word-finally (Bergsland 1986: 98). This is not dramatically different from Central Yup'ik, the Eskimo-Aleut language in the sample, so that diachronic phonological pressure indeed does not seem to play a major role.

"Proto-[Eskimo-Aleut] must have lost a large portion of its previous stock of lexical items as capitalisation on its highly productive derivational apparatus increased and lexical gaps were filled more and more by derived forms from relatively few stems." Indeed, there are some 200 basic postbases (see § 4.4.2. for this term) in both Greenlandic and Central Yup'ik, with somewhat fewer in Aleut; 50 Aleut postbases have Eskimo cognates (Bergsland 1986: 102), suggesting an expansion of derivational postbases in the latter. Fortescue (1992: 245) argues that an expanding derivational apparatus with concomitantly increasing allomorphy creates an increase in memory load for the acquisition of its properties, and, once the process has begun, feeds into the development of polysynthesis, the outcome of which is a "reorganized state of balance between the inventory of lexical stems as opposed to productive bound affixes." Furthermore, he (1992: 246) states that "typically, polysynthetic languages do display a relative paucity of lexical stems, this being counterbalanced by an enormously increased derivational potential compared to more analytic languages." It would indeed be of great value to assess this impressionistic statement quantitatively on the basis of a sample of languages with a high degree of synthesis. But even with more systematic evidence pending, Fortescue's account has some merit in that it could explain a high degree of analyzable terms in many "polysynthetic" languages, in spite of a universally accepted definition still lacking. Note, however, that languages regarded as polysynthetic are not necessarily characterized by a rich derivational apparatus, Ket being an example of such a language (cf. § 4.5.2.1.). Furthermore, Fortescue's proposal would also fail to account for the behavior of languages with an isolating profile, such as Efik, Bororo, and Hawaiian. This should not be taken to mean that his proposal as to a shrunk inventory of lexical elements at the expense of increasing derivational possibilities is incorrect, but merely that it cannot account for all cross-linguistic variation with respect to differences in the percentages of analyzability, since, if indeed syntheticity were the sole responsible parameter, one would expect only such languages with a profile of lexicon-grammar-interaction as outlined by Fortescue for Eskimo-Aleut to be characterized by a largely analyzable lexicon, and not others. The following section discusses grammatical properties as a potential factor, with particular reference to a typical ingredient of polysynthesis: head-marking.

5.4.2.12.5. Other Grammatical Factors in the Distribution of Morphological Complexity?

Thanks to the work of Nichols (e.g. Nichols 1992, 1998, Bickel and Nichols 2009, in press), it is well-known that there is a world-wide cline in the distribution of certain grammatical features, such as head- vs. depending marking, inclusive/exclusive distinction in pronouns, numeral classifiers, as well as consonantism in pronominal roots (Nichols and Peterson 1996).

Could it be possible that there may also be grammatical factors that shape a language's behavior with respect to analyzability in its lexicon? Consider, for instance, the following examples of basic transitive constructions from Kiowa and Biloxi, which are typical for languages of North America.

(6.) a. *kɬyą́·hî· tʰàlí· ę́-góp*
man boy 3SG/AGT:DU/OBJ-hit/PF[44]
'The man hit the two boys' (Watkins 1984: 205)

b. *tohoxka ayeki duti na*
horse corn he.eats.it
'The horse eats the corn' (Einaudi 1974: 166)

These are head-marking constructions (Nichols 1986): arguments carry no markers indicating their grammatical function in the clause; rather, these are identified by means of affixes on the verb. In contrast, languages of Eurasia are predominantly dependent marking, as illustrated by the sample languages Bezhta:

(7.) *gedi ãq'o boxx-iyo*
cat.ERG mouse catch-PST.W[45]
'the cat caught the mouse' (adapted from Xalilov 1995: 410)

However, not all languages of Eurasia follow the typically dependent-marking clause alignment in this area. Two notable exceptions are the Yeniseian and Munda language families, as illustrated by Ket and Sora examples in (8.).

(8.) a. *hīɣ qímdɨ̈l dítòŋ*
hīɣ qímdɨ̈l du^{8}-i^{6}-t^{5}-a^{4}-oŋ^{0}
man girl 3M.SJ^{8}-3F.O^{6}-SU^{5}-D^{4}-see^{0}[46]
'The man sees the girl' (Vajda 2004b: 22)

b. *ənlen daʔa-n a- tiy- t- ay*
we water-N.SFX 1PL-give-NPST-1[47]
'We give (him/her) water' (Anderson and Harrison 2008: 328)

Sora in fact features the object marker *a'dəŋ* occurring in connection with lexical rather than pronominal arguments. It is grammaticalized from the possessed form of a word meaning 'body,' and is probably a recent innovation, as suggested by the fact that it is an independent word rather than an affix and restricted to animates, a semantic restriction that is typical for early stages in the grammaticalization of case markers (Hopper and Traugott 2002). Other major Munda languages, e.g. Mundari and Santali, lack marking of core arguments altogether, making Sora an unusual Munda language in this respect.

[44] Glosses: AGT 'agent,' DU 'dual,' OBJ 'object,' PF 'perfective.'
[45] additional gloss: PST.W 'witnessed past.'
[46] glosses: M 'masculine class (a subset of animate class),' SJ 'verb-internal subject agreement affix, or subject pronoun,' F 'feminine class (a subset of animate class), O 'verb-internal direct object agreement affix, or direct-object pronoun,' SU 'suppressive adposition (verb affix denoting superficial contact with an object), D 'durative marker (appears in many stative and activity verbs).'
[47] additional gloss: NPST 'non-past.'

Crucially, not all languages of the Americas are head-marking. One example of a strictly dependent-marking North American language is Wappo, formerly spoken in California (Wappo is a so-called marked nominative language, but this does not affect its characterization as being dependent marking), as seen in (9.).

(9.) *ce k'ew-i ce holo:wik'a t'a-ta?*
DEM man-NOM DEM snake kill-PST[48]
'the man killed the snake' (Thompson et al. 2006: 11)

The point of discussing these examples is that Ket and Sora, typologically unusual languages for Eurasia, receive after Abzakh Adyghe (which is a double-marking language on the level of the clause in terms of Nichols 1986, but has many head-marking traits) the highest scores in the degree of analyzability in the nominal lexicon, while Wappo, a typologically unusual language for North America overall, receives the lowest score in analyzability of all North American languages in the statistics samples. This suggests that there are other structural features, aside from phonology, in play when it comes to the shape of the nominal lexicon. In § 4.6.5.4., it was suggested that head-marking elements in Kiliwa may be a factor facilitating the coinage and conventionalization of complex clausal nominals. While head- as opposed to dependent-marking is a typological factor that may be applied on different levels of linguistic structure, including morphological marking within the noun phrase as well as clause-level and even interclausal syntax (and these patterns may be used to jointly define a profile of individual languages with respect to the parameter, Nichols and Bickel 2005b), the focus will here be on the level of the clause, at the expense in particular of marking in the noun phrase. This is not to say that NP-level marking would not be interesting to investigate.

The question whether there are differences in the lexicon depending on preferred marking patterns on the clause level is tested in the following fashion: rather than using one overall metric assigning languages to one type (Nichols and Bickel 2005a), the data on verbal person marking from Siewierska (2005), with amendments to fill gaps in the data as already used in § 4.7., provides one measure of indexing on the verb (=head-marking). In addition, data were gathered for the sample languages on whether core grammatical relations are flagged by case markers or case-like elements such as adpositions (=dependent-marking). Data are in Appendix C. This in effect creates two independent parameters for head- and depending marking elements (cf. Cysouw 2002): a dependent-marking language on the clause level is defined as one with core cases but no verbal person marking, while a language with dominant head-marking elements on the clause level is one with no core cases, but verbal person marking for both A and P arguments. As tables 19 and 20 show, the properties are areally unevenly distributed under the Dryer-6 breakdown (cf. also Nichols 1992). The differences are significant at $p < .001$ for verbal person marking and at $p < .02$ for presence vs. absence of core cases by Fisher's exact tests, so statistical modeling once again needs to take these differences into account.

[48] Glosses: DEM 'demonstrative,' NOM 'nominative case,' PST 'past tense.'

	Africa	Australia-New Guinea	Eurasia	North America	South America	Southeast Asia and Oceania
No verbal person marking	2	2	3	1	3	5
Only A	1	1	4	1	4	0
A and P	0	2	4	15	8	0

table 19: areal breakdown of types of verbal person marking

	Africa	Australia-New Guinea	Eurasia	North America	South America	Southeast Asia and Oceania
Languages with core cases	2	7	9	5	12	1
Languages without core cases	2	3	2	16	13	5

table 20: areal breakdown of presence vs. absence of case marking for core grammatical relations

Contrary to the hypothesis generated by looking at the languages mentioned above, no clear impact of differences in the locus of marking on the clause-level was revealed by a Mixed Model controlling for area emerged, neither for the combination of the two variables of verbal person marking and core cases (p = .4476), nor for one of them separately (person marking: p = .6484 , core cases: p = .3134).

As observed by Nichols and Bickel (2005b), genetically related languages sometimes differ in their marking type. In particular, they note that within Uto-Aztecan, Pipil is a consistently head-marking language, without cases for core grammatical relations but with affixes on the verb cross-referencing arguments (Campbell 1985: 39-56; 74), while Tümpisa Shoshone is consistently dependent marking, featuring a nominative-accusative case system and no indexing of arguments on the verb (Dayley 1989a: 53-54; 176-178).

To see whether there is any noticeable impact of these differences, Tümpisa Shoshone equivalents for the full 160-meaning list were gathered from Dayley (1989b), yielding 126 of 160 possible equivalents (this was done after most calculations were computed and data for chapter 6 were collected, so the Tümpisa Shoshone data is not otherwise evaluated systematically). The result is that 29 per cent of these were analyzable (as opposed to 18.2 in Pipil), with 34.3 per cent being of the derived type (as opposed to 25.9 per cent in Pipil). According to the hypothesis, Pipil should have a higher number of analyzable terms than Tümpisa Shoshone, but it does not, thus showing that any immediately effects of locus of marking in the clause on the degree of analyzability in the lexicon seems unlikely.

5.4.2.12.6. Effects of Mode of Subsistence on Analyzability in the Lexicon? The lexicon is probably the subsystem of language which is most directly influencable by non-linguistic factors, be they cultural or environmental. It is therefore conceivable that the lifestyle of a speech community will be a factor that influences the structure of language, as suggested e.g. in Brown (2005a, b) for certain features of the lexicon specifically (see also Cysouw and Comrie forthcoming for some possible grammatical correlates). Two different data sources are used to address whether there are such differences in the major quantitative variables concerning the lexicon surveyed in this work. Hammarström (2010, online appendix) provides data on the dominant mode of subsistence for the world's language families. Hammarström employs a binary classification into hunter-gatherers and agriculturalists. In addition, data from Murdock and White (1969) provide more detailed information on mode of subsistence for a selection of world cultures. On the basis of the information from Murdock, languages where grouped according to whether the main contribution to mode of subsistence is provided by (i) hunting and gathering, (ii) horticulturalism or pastoralism, or (iii) advanced agriculture. In cases where two of the above factors are said to contribute equally, the culture and its corresponding language was coded as belonging to the category with the lower number. For instance, cultures which rely on both hunting and gathering and horticulturalism or pastoralism were treated as hunter-gatherers. This is simply a measure to avoid ambiguities and thus to allow for statistical analysis. However, the overlap between them and the corresponding languages presently surveyed is rather small, which is why the data from Murdock (1969) were amended by extraction of relevant information from Levinson (1991).[49] Resulting data are in Appendix C.

Testing for effects on the degree of morphologically complex terms on the basis of both datasets for mode of subsistence using a Mixed Model controlling for areal effects reveals at best borderline significance for the Hammarström dataset (likelihood ratio test: *p*-value for factor: = .1010, estimate: 3.743) and no significance for the Murdock and White/Levinson dataset (likelihood ratio test: *p*-value for factor =.2335). Figures 35 and 36 plot the results.

[49] When data on a particular group are available in both sources, data from Murdock (1969) were used.

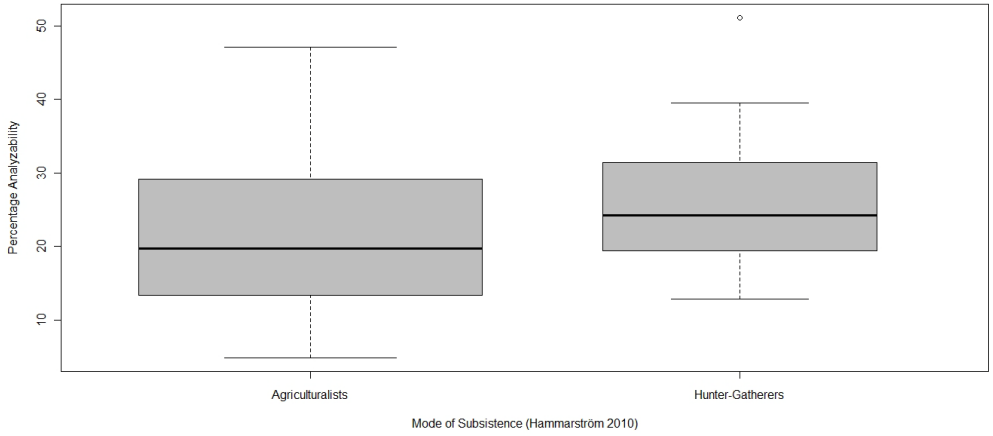


fig. 35: differences in the degree of analyzable terms between agricultural ists and hunter-gatherers, data from Hammarström (2010)

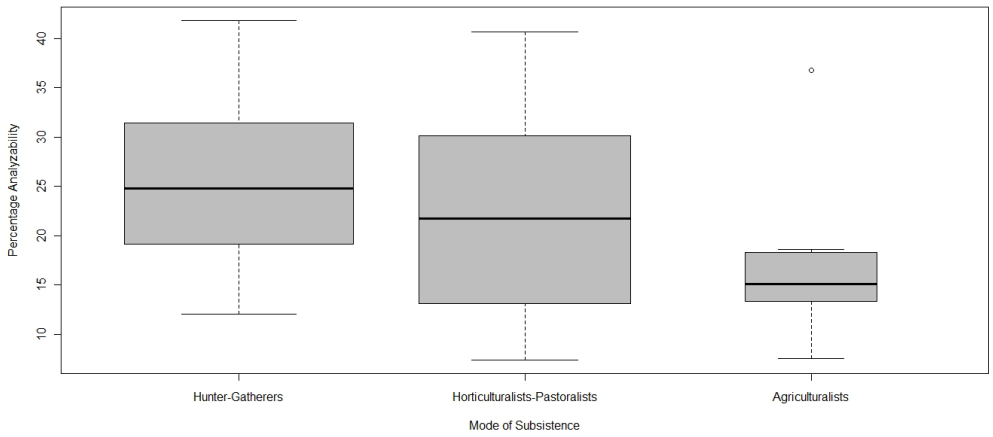


fig. 36: differences in the degree of analyzable terms between agriculturalists, horticulturalists/pastoralists and hunter-gatherers, data from Murdock and White (1969) amended by data from Levinson (1991)

In both cases, languages spoken by peoples relying dominantly on hunting and gathering as the primary mode of subsistence turn out to employ morphologically complex expression to a higher degree than agriculturalists, although there also is an areal signal in the data, in particular in the Murdock and White/Levinson data, and more importantly, the difference is insignificant. If there should be a genuine effect in spite of lack of clear statistical significance, it points to a further area of interaction between language and culture

which seems worth probing in more detail, although at present it is not clear just what should be the cause of the effect of mode of subsistence on the structure of the lexicon. Like Cysouw and Comrie (forthcoming), here, the discussion abstains from speculating about possible causes and restricts itself to simply reporting what can be observed.[50]

5.4.2.12.7. Creolization. One notable property of the behavior of the two creoles in the sample, Bislama and St. Lucian Creole French, is that it is remarkably unremarkable. This is in stark contrast with the commonly uttered opinion that "[p]idgins and creoles exhibit a high degree of motivation and transparency in compounding as a direct consequence of their small vocabulary" (Romaine 2002: 1094).[51] The evidence gathered for this study receives additional backup by the fact that the Creole language in the sample of Haspelmath and Tadmor (2009c), Seychelles Creole, receives the second lowest simplicity score of all languages in the sample exceeded in rarity of complex expressions only by Tarifiyt Berber (Bradley Taylor p.c.). If there is anything remarkable about the behavior of the creoles in these two studies, it is that they, quite contrary to what one might expect, come out rather at the lower end of the continuum on which languages are categorized with respect to the presence of analyzable terms. There are, judging from the evidence from both samples, a large number of non-creoles that regularly outreach the creoles with respect to the quantity of complex terms. This might be interpreted as being due to the fact that creoles, in the process of evolution from earlier pidgins (if they indeed evolved from pidgins, the notion that this always needs to be the case seems to be increasingly questioned), lexicalization (in the diachronic sense of univerbation) has rendered a large number of erstwhile compounds or circumlocutory phrases unanalyzable (a process alluded to by Romaine 1988, 2002: 1094 for Tok Pisin specifically). However, there is no evidence for such a development on a large scale in the data for Bislama and St. Lucia Creole French. Rather, the simplex lexical items in each language can be readily traced back to simplex lexical items in the respective lexifier language.

Likewise, referential expansion by means of polysemy as a technique to enrich expressive possibilities in creoles, as mentioned e.g. by Holm (1988: 108), is not present to a significantly higher degree when compared with the world-wide situation in non-creoles. On the basis of this study, it is not possible to confirm or refute claims such as that "there

[50] A further possible correlation that was tested is that between the occurrence of colexification of 'milk' and 'breast.' When one thinks about what pragmatic factors may give rise or maintain the colexification of the two referents, it seems obvious that it must be utterances in the context of nursing, such as "the baby wants [milk/breast]," where in fact reference is ambiguous (compare the following example sentence for Cashinahua *chuchu*: *Chuchu manuikiki. Amave.* 'He wants milk/breast. Let him drink.' (original translation: "Desea leche. Hazle tomar."). In contrast, in societies with an advanced mode of subsistence involving domestication of animals, in other words, where milk may be assumed to be a regular part of the diet, other contexts in which 'milk' occurs may in fact be more salient. However, the results were negative: there was no effect of mode of subsistence on this pattern of colexification when testing with both datasets on mode of subsistence.

[51] Relevant for the present study in general is also Rice's (to appear) description of the inventory of lexical stems in Athapascan languages as "staggeringly small," which is why according to her stems are "routinely called upon semantically to do double and triple duty, if not more, through conversion, compounding, juxtaposition, and inflection." Note that the analyzability score for the sampled Athapascan language Carrier is among the highest in Western North America.

is ... every indication that the lexicons of early (i.e. non-extended) pidgins are very much smaller than those of natural languages" (Holm 1988: 108), simply because the two languages in the sample are creoles and not pidgins in their early state of their development. But as far as creoles are concerned, from a cross-linguistic point of view, there is nothing special to be noted about them. However, statements emphasizing a high degree of motivation in pidgins and creoles do have some justification. When compared with their lexifier language, it may well be true that there indeed is a notable increase in the usage of compounds and polysemy, as can be illustrated by contrasting Bislama and English: a number of unmotivated simplex terms of English have not made their way into Bislama, their meanings being rendered by complex items, such as *ashes* vs. *sit blong faia* 'shit of fire,' *nest* vs. *bed blong pijin* 'bed of bird,' and many more. Likewise, *smok* in Bislama can not only mean 'smoke,' but also 'dust,' and *nus* does not only denote the 'nose' but also 'nasal mucus' and 'froth, foam.' But note that English, although not included in this study, is extremely likely to participate in the language area comprising Eurasia and Northern Africa with low degrees of complex lexical items and polysemes (evidence for the validity of this assumption is in Urban 2008). So while the degree of lexical motivation is elevated in creoles when compared to the largely unanalyzable stock of vocabulary items in the lexifier languages, which is, at least in the case of Indo-European-based pidgins and creoles unusually poor in motivated words, there seems to be no basis for the claim that creoles in general have elevated ratios of complex and polysemous lexemes when compared against the cross-linguistic situation. Rather, they seem to have, like their European lexifier languages, a comparably low degree of motivated terms, although somewhat higher than the languages they are descendent from.

As far as the specific semantic associations by means of compounding and the types of occurring semantic extensions of lexifier-language lexemes are concerned, substrate influence rather than creolization-specific universal processes appear to plays a significant role. For instance, the presence of a number of complex expressions on the basis of *sit* 'shit' in Bislama, one of which was mentioned above, appears to mimic structurally similar formations that seem to be common in Oceania as a whole (see § 6.2.3.3. for discussion of such extensions). For the case of Bislama specifically, Camden (1979) amasses evidence that the semantic structures in the lexicon (as well as in syntax) in particular match that of the Oceanic language Tangoa to a high degree, his conclusion being that "while the Bislama lexical structure looks basically English to a native speaker of English, it also looks basically Tangoan to a native speaker of Tangoan" (1979: 54). Similarly, semantic extensions of lexical items in Jamaican creole noted by Cassidy (1971: 216), such as the extension of the word for 'sun' to also mean 'day' and the ability of the word to 'water' to also refer to bodies of water such as 'river' or 'lake,' is frequent in normally transmitted languages globally, including African languages that form the substratum of Jamaican Creole. Holm (2000: 104), in discussing compounds in Nubi, an Arabic-based creole of Africa, mentioned by Heine (1982: 20), notes that "[s]uch compounds may have resulted from a universal strategy for expanding a pidgin vocabulary to fill lexical gaps, or they could represent calques on compounds in substrate languages." Similar evidence is presented in Parkvall (2000: 113-114), leading him to assume an agnostic position as to the source of lexical structures in Atlantic creoles as well. While the more or less anecdotal evidence

presented above does not rule out the possibility that there may indeed be mechanisms of lexical expansion by formation of morphologically complex expressions peculiar to the process of creolization only, nor that it may indeed have happened that semantic extensions occurred in the context of attempting communication in a setting with extremely little shared vocabulary between interlocutors, in the light of the ubiquity of most semantic structures found in creoles (and sometimes thought to be peculiar to creoles), substratum influence seems in many cases to provide a simpler and more parsimonious explanation for semantic structures in creoles (see already Huttar 1975, who arrives at similar conclusions, albeit on a somewhat different route). At any rate, the data in chapter 6 may be of use for creolists in formulating more fine-grained hypotheses as to the question of the origin of creole semantic structures.

5.4.2.12.8. Concluding remarks. Previous sections discussed alternatives to an explanation in terms of phonological complexity and root structure. It turned out that, although effects of some of them cannot be ruled out, accounts based on them would be less stringent than the one appealing to complexity of the word and of the sound system, either because (i) the phenomena in question are not universally applicable since they pertain to certain types of languages only, or (ii) mostly yielded negative or equivocal results when analyzed by means of statistics. In summary, structural pressure arising from complexity of the word and of the sound system can for the time being be said to be the most plausible candidate to shape analyzability in the lexicon (although further evidence not presently available on each of the topics may change this assessment), even though the mechanisms underlying it are not entirely clear in their details, and the particular interpretation suggested here is open to revision and refinement, with studies of homonymy in actual speech events being sparse as they are.

At any rate, the obtained correlations remain an empirical fact. As Dryer (2003: 120) remarks: "While I share the interest that others have in explaining crosslinguistic generalizations, there is a sense in which such generalizations are more valuable than the hypothesized explanations, since we can often have a much greater degree of confidence in the validity of the generalizations themselves than we can have in the explanations that have been hypothesized for them." Then, the discussion can be concluded with a dialogue from Orr (1962: 17) that seems appropriate:

R. Do you mind if I rest awhile? I've just found a monster floating about in my psychological orbit, and I feel a little uneasy.
O. What is it?
R. Synonymic-homonymics, an ugly brute!
O. Perhaps we had better stop for a bit.

As another alternative, perhaps, as suggested by David Gil (p.c.), it would also be worth thinking in the opposite direction: if languages favor complex terms, they can live with simple phonological inventories (cf. also the alternative explanation, though not generally accepted, for the developments in Mandarin Chinese discussed in § 5.4.2.3.2., which has it that the phonological system only began to shrink after the introduction of disyllabicity).

Instead of elaborating on these issues any further, the following discussion is concerned with the focussing on nominal referring expression in this study, seeking at least to hint at some interdependencies with the overall lexical organization of the language's lexicon in terms of the two parts of speech held by most linguists to be universal: nouns and verbs.

5.5. NOUNS AND VERBS

5.5.1. General Analysis of Vocabulary

Having established that there are languages in which analyzability is pervasive in the nominal lexicon, a question one can ask is whether this is due to a general difference in the prevalence of nominal as opposed to verbal encoding of referents. Relatedly, the question also pertains to the relative frequency of simplex noun and verbs in the lexicon, which may be relevant, because, if a paucity of simple unanalyzable nouns can be diagnosed for a particular language, then this would correlate with an elevated degree of analyzable nouns as a sort of "replacement vocabulary" to make up for the paucity of root nouns. To contextualize the investigation, it should be pointed out that highly divergent organizations of the nominal and verbal domain have been noted in the literature. Pawley (1993) reports that in Kalam, a language of New Guinea, verbs are a closed class with very few members and quite generic semantics which are conventionally combined in larger constructions to yield more specific semantic content, while nouns are in contrast much more numerous, although also here morphologically complex expressions are found. In contrast, Talmy (2000) highlights the deverbal character of the Atsugewi nominal lexicon. While Kalam thus makes do with a small restricted set of verbs, they are of such importance in Atsugewi that they are the basis for the formation of the other major part of speech, nouns. In this sense, the investigation takes up the rough typology of basic lexical types (i.e. noun-based vs. verb-based) outlined by Talmy (2000), which is with a different approach also addressed by Nichols and Nichols (2007).

Elucidating this question is not easy since what is needed is a representative sample of general vocabulary for all languages to be tested (Nichols and Nichols 2007 restrict themselves to a small list of glosses the equivalents of which they search for in their test languages of the Caucasus and the Pueblo languages of North America). Since a more general assessment that aims at looking at the vocabulary as a whole is very time-consuming, the present investigation is restricted to a small set of test cases, consisting of data for only four languages, and because of this restriction, the generalizations to be drawn can be nothing but extremely tentative.

Representative languages were selected more or less at random, except for the fact that obviously they were meant to define extreme points on the continuum of analyzability in the nominal domain. Since what is of interest here is the behavior of languages with a highly analyzable nominal vocabulary, two such languages, Kiliwa and Pawnee, were analyzed to allow for comparison. Badaga was chosen as a language representing the opposite type, with very few analyzable nouns, and Koyraboro Senni as a language that falls somewhere in between the extremes. An important criterion was that dictionaries

are sufficiently large and can thus assumed to be more or less comprehensive. Another important requirement for the selection of languages was that the consulted source provide clear information as to the part of speech of the headwords; another criterion was that there is a grammar available (written in the case of Koyraboro Senni, Pawnee, and Kiliwa by the same author as the lexical source) that identifies the morphosyntactic criteria that allow for distinguishing between nouns and verbs (Parks 1976, Mixco 1965, 2000, Heath 1999, Balakrishnan 1999). The methodology is simple: a random sample of the vocabulary was gathered by reading every tenth page of the Pawnee/Kiliwa/Koyraboro Senni-English section of the dictionaries, and, due to its larger size, every 20th page of the Badaga-English section of the dictionary, beginning on the first page (see Nettle 1995 for a similar approach for generating a random vocabulary sample from dictionaries). This avoids both biases from (fossilized) prefixes of a certain shape that cause a particular part of speech to begin with a certain segment and thus to cluster in a certain region of the dictionaries (note that, for instance, in Meyah, many nouns begin with /m/, Gravelle 2004: 104). Entries for native unanalyzable nouns and verbs (that is, disregarding clear loanwords) on the pages read were counted and, in the case of nouns, their meanings were recorded alongside. In Koyraboro Senni, many stems are ambiguous as to lexical category and can function as either nouns or verbs (Heath 1999: 96). Such stems were not counted as being either nominal or verbal and were simply ignored. The same goes for the fewer number of such cases in the other languages, such as Pawnee stems functioning as verbs but that may be used as nouns by suffixation of the nominal suffix *-uʔ*.

This yielded a sample of 101 Pawnee words, 68 Kiliwa words, 177 Koyraboro Senni words, and 145 Badaga words, coded for whether they are defined by language-internal criteria as nouns or verbs. Subsequently the number of nouns was divided by the number of verbs to obtain a measure called the NOUN/VERB-RATIO here. A high noun/verb-ratio indicates that simplex nouns are more frequent than simplex verbs, and a low ratio indicates the opposite situation: unanalyzable verbs outnumber unanalyzable nouns. Table 21 provides the values for the noun/verb-ratio from the dictionary sample along with the number of analyzable terms on the list of 160 meanings.

	Percentage Analyzability	Noun/Verb-ratio
Badaga	9.4	7.056
Koyraboro Senni	13.6	1.77
Pawnee	47.1	0.46
Kiliwa	51.1	0.66

table 21: noun/verb-ratio and percentage of analyzable terms for the four test languages

As the values already show, the language with the highest noun/verb-ratio and therefore with the largest number of unanalyzable nouns, Badaga, is also the one with the fewest analyzable terms on the 160-meaning list, while the two North American languages with pervasive analyzability in the nominal lexicon, have a very low number of simple nouns as opposed to verbs (cf. in this context also discussion of Kiliwa summarized from Mixco 1965, 2000 in § 4.6.4.2.1., where the "verbal" character of the language is emphasized).

The Spearman's rank correlation is very strong at -0.8, and it is easy to mentally fit a regression line. Although the results must be seen as being preliminary in nature, they cast doubt on Dixon's (2010: 305) as strong as casual claim that "[t]here are never as many simple verbs as there are nouns." A plot of the correlation is in figure 37.

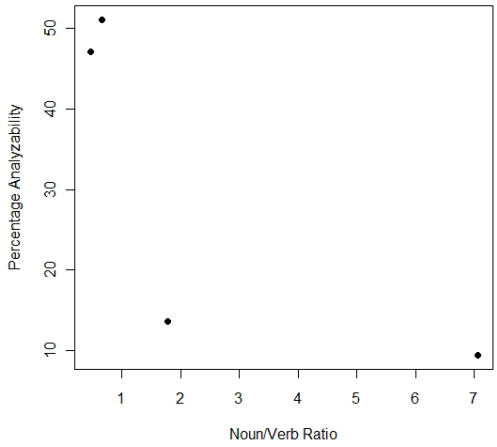


fig. 37: correlation between the noun/verb ratio and the percentage of analyzability

Another question that can be addressed with this data is: if the inventory of complex nominals in a language is very large and covers many meanings, what meanings, then, are expressed by simple nouns? This yields quite interesting results. As seen in table 22, unanalyzable nouns in Pawnee sampled from the dictionary are easily assigned to a small number of semantic domains: somewhat less than half of them are terms for animals and plants on the generic level. Other semantic domains in which unanalyzable nouns are found are kinship terms, body-part terms, topological and natural kind terms, and, frequently, names of tribes and ethnic or social groups. Terms for artifacts are not on the list. Similar results, in particular absence of simplex artifact terms, are found by Nichols (2008) for Zuñi.

Domain	Number	Percentage	Example
(i) flora and fauna	13	41.94	*akiwaasas* 'black haw'
(ii) kinship	4	12.90	*-kaa-* 'grandmother'
(iii) body-parts/body-related meanings	2	6.45	*iit-* 'body, corpse, carcass'
(iv) nature-related/topology	3	9.68	*huupirit* 'star'
(v) tribes, ethnic or social groups	6	19.35	*Pasaasi* 'Osage Tribe, Osage Male'
(vi) professions/special persons	1	3.23	*Ctuʔuʔ* 'Witch Woman, a mythological old woman who has supernatural power'
other	2	6.45	*awi* 'fleeting image; quick motion'
total	31		

table 22: Pawnee simplex nouns in the sample according to semantic domain

The analysis for Kiliwa yields quite similar results, summarized in table 23. Here, too, a little less than fifty percent of sampled simplex nouns are flora and fauna terms. In contrast to Pawnee, no kinship terms are found on the sampled pages, but body-parts (somewhat more than in Pawnee), nature-related meanings, and one name of a tribe figure on the list. The noun/verb-ratio in Kiliwa is also quite low, but somewhat larger. Correspondingly, some Kiliwa nouns fall in semantic domains not attested for Pawnee in the sample: there are two native simplex nouns for artifacts, and there is one term denoting an abstract property presumably applicable to many entities.

Domain	Number	Percentage	Example
(i) flora and fauna	14	51.85	*nxil* 'Pitahaya'
(ii) kinship	0	0	-
(iii) body-parts/body-related meanings	6	22.22	*-ha?* 'mouth, voice, breath'
(iv) nature-related/topology	1	3.704	*-kwiy* 'cloud'
(v) tribes, ethnic or social groups	1	3.70	*xwa* 'warrior; enemy, foreigner; principally Cocopa'
(vii) artifacts	2	7.41	*cpat* 'door'
(viii) abstract relations/properties	1	3.70	*cpa?* 'proejction, protrusion, end, tip'
(ix) culture/mode of subsistence/food	2	7.41	*'kuskuwpl* 'edible grass seeds'
total	27		

table 23: Kiliwa simplex nouns in the sample according to semantic domain

Nichols (2008) argues that there are lexico-semantic restrictions in Zuñi as to what a simple noun may denote. In particular, according to her analysis, they are constrained to natural kinds, that is, excluding artifacts. She proposes that this is the explanation for the extremely few loanwords in Zuñi. Given the preliminary results obtained here, this may be true of other North American languages as well, though probably not of all, as Nichols (2008), drawing on data from Brown (1999) also notes that noun borrowability in other Pueblo languages is less constrained (cf. also § 5.4.2.7.1. on borrowing in an American context). Leaving the Americas for Africa to investigate the semantics of simplex nouns in Koyraboro Senni, drastic differences are immediately noticeable. Koyraboro Senni also features many simplex nouns for animals and plants, although the percentage is somewhat depressed when compared with the American data. It has a comparable portion of simplex nouns in the domains of kinship, body-parts, and nature-related terms. From this does not follow that Koyraboro Senni has fewer monomorphemic nouns for animals and plants in absolute numbers, but rather, that their relative percentage is depressed by the presence of monomorphemic terms in other semantic domains. This is noticeable in the domain of artifacts, but particularly obvious in the emergence of terms related to culture, mode of subsistence (the speakers are pastoralists and agriculturalists), and for social relations. Likewise, a term for 'bird' is among the recorded meanings (although it would be wrong to conclude that this is due to increase in societal complexity, since Pawnee also has a simplex noun for 'bird' not on one of the sampled pages). Furthermore, there is a noticeable rise in simplex nouns for abstract relations, properties and quantities. Among

semantic domains of simplex nouns found neither in Pawnee and Kiliwa are those of temporal concepts, such as phases of the day and seasons as well as one noun to denote the emotion 'anger.'[52] Table 25 provides a summary of the Koyraboro Senni data.

Domain	Number	Percentage	Example
(i) flora and fauna	43	36.76	*addihijji* 'aardvark'
(ii) kinship	3	2.56	*feŋge* 'sibling-in-law'
(iii) body-parts/body-related meanings	12	10.26	*diini* 'gums'
(iv) nature-related/topology	8	6.84	*karji* 'thorn, barb'
(v) tribes, ethnic or social groups	4	3.42	*sače* 'ethnic group specializing in leather amulets'
(vi) professions/special persons	1	0.85	*gariibu* 'beggar'
(vii) artifacts	13	11.11	*ferow* 'brick'
(viii) abstract relations/properties	4	3.42	*baka* 'handful'
(ix) culture/mode of subsistence/food	17	14.53	*herow ~ herew* 'young nanny-goat (not yet a mother)'
(x) life-form terms	1	0.85	*subu* 'grass, herb'
(xi) social relations/business	5	4.27	*yaahi* 'friend, pal'
(xii) place names	1	0.85	*bamakoo* 'Bamako'
(xiii) emotions	1	0.85	*zattu* 'desire (for sth.)'
(xiv) temporal concepts	3	2.56	*lahula* 'winter, cold season'
other	1	0.85	*baali* 'pulp (of fruit)'
total	117		

table 24: Koyraboro Senni simplex nouns in the sample according to semantic domain

This trend is continued in Badaga, the language with the highest noun to verb ratio. As table 25 shows, the ratio of flora and fauna terms is further depressed, while the domains of kinship, body-parts and nature-related meanings are relatively constant in their percentages across languages, and the domains of artifacts and abstract relations and properties are represented to about equal percentages in Koyraboro Senni and Badaga. In Badaga, however, there is a dramatic increase in terms having to do with social and religious organization that may be due to an increasingly complex social organization and social stratification. While there are, unlike Koyraboro Senni, no recorded instances of nouns encoding temporal concepts (although they surely must exist), there are many more emotion terms that are encoded nominally rather than verbally in this language, and an additional semantic domain of simplex Badaga nouns not found in the languages discussed so far are units of measurements.

[52] An informal browse through Park and Pratt (2008) reveals that emotions are indeed encoded in Pawnee mostly by verbs, while there are basic nouns for temporal concepts.

Domain	Number	Percentage	Example
(i) flora and fauna	15	11.54	*mundari* 'vine'
(ii) kinship	5	3.85	*auve ~ avve* 'mother, father's wife, wife's father's sister; Toreya term of address for higher-status Badaga women'
(iii) body-parts/body-related meanings	8	6.15	*moḷḷe* 'navel, male nickname'
(iv) nature-related/topology	13	10	*ailu* 'dewdrops, beads of dew'
(v) tribes, ethnic or social groups	2	1.54	*Bekkan* 'Bekkan, Pekkan ...'
(vi) professions/special persons	15	11.54	*haika* 'unintelligent man; male nickname,' 'horseman, equestrian, cavalier; male name'
(vii) artifacts	17	13.08	*moḷe* 'nail, peg, branch'
(viii) abstract relations/properties	4	3.08	*haetu ~ aetu* 'old things'
(ix) culture/mode of subsistence/food	14	10.77	*ha:yi ~ ha:i* 'farmland near a village'
(x) life-form terms	1	0.77	*hakki ~ akki ~ akkilu ~ hakkilu* 'bird, avifauna'
(xi) social relations/business	15	11.54	*saṇḍe* 'war, fight, quarrel'
(xii) place names	2	1.54	*Cocci* 'Cochin ...'
(xiii) emotions	3	2.31	*a:ti* 'wreath; cyclical movement, circular motion, ritual offering'
(xv) units of measurement	2	1.54	*aigua* 'five measures (ca 18.53 litres)'
(xvi) theology	4	3.08	*de:varu* 'god, gods, deity'[53]
other	10	7.69	*saḍunga* 'jingle, jingling sound'
total	130		

table 25: Badaga simplex nouns in the sample according to semantic domain

It has been, intuitively plausibly, claimed that size of vocabulary increases with technological evolution (Witkowski and Burris 1981), and this is congruent given the expansion of specialized cultural vocabulary in Koyraboro Senni and Badaga. However, the methodology Witkowski and Burris employed is dubious: they simply take dictionaries for a number of languages counting the number of entries, and find languages spoken by large industrialized speech communities to have more entries, concluding that "large-scale societies have larger lexicons than small-scale societies" (1981: 144). They acknowledge that dictionary size depends on purpose, but ignore the issue of comprehensiveness and the very different circumstances under which dictionaries for "large" and "small" languages are

[53] This is in fact the plural of *de:va* 'god, godling, deity,' but has its own entry (*de:va* is often used as a honorative singular).

typically created. Specialist vocabulary is said to increase, while 'core' vocabulary remains constant in size. Names for specific plants and animals are said to decrease, and this is consistent with the Badaga results, the one of the investigated languages spoken in the most socially developed speech community.

Summing up, the preliminary evidence from the investigation is that where unanalyzable nouns are few in number, most of them are names of specific animals and plants, with some additional ones in the domains of kinship, nature-related terms, and sometimes artifacts. In languages where they are more frequent, they also cover culture-related meanings (with "culture" perceived in the broadest possible sense), and extend more frequently to also denote abstract concepts as well as emotions. In this context, note that names for animals and kinship terms are precisely the meanings for which unanalyzable basic nouns can be reconstructed for Indo-European, a language in which the nominal lexicon appears to have been characterized by analyzability to a high degree (see § 5.4.2.7.2.).

5.5.2. VERBAL VS. NOMINAL ORIENTATION OF BODY LIQUID AND AEROSOL TERMS

Another aspect of differing lexical organization in terms of nouns and verbs comes from the semantic fields of body liquids and aerosols (as used in physics, i.e. smoke, steam, fog, clouds). In the majority of sampled languages, meanings in both domains are encoded lexically as nouns. However, at times, the morphologically basic expression which encodes them are verbs, not nouns. This is also true of some other meanings in the database. For instance, like a number of other languages in the sample, Ineseño Chumash has a term for 'belt,' *qanatiʹš*, which is derived from the verb *qanatiʹ-* 'to put on a belt.' But differences in the domain of body liquids and aerosols are worth looking at in more detail because they are semantically well-circumscribed, and it is here that differences in lexical organization are most eye-catching. A language in which many attested terms in these domains are basically verbs or derived from other non-nouns is Nuuchahnulth, with the corresponding noun derived from them by the nominalizer *-mis* (which also occurs as a free-standing noun 'thing') or other derivational suffixes:

(10.) a. 'cloud': *ɬiẇaḥmis* /ɬiẇaḥak-mis/ 'be.cloudy-NMLZ'
b. 'fog': *ʔučqmis* /ʔučqak-mis/ 'foggy-NMLZ'
c. 'smoke': *qʷiš-aa* 'to.smoke-??'
d. 'steam': *muqckʷii* /muq-ckʷɨ̌/ 'to.steam-remains.of'
e. 'blood': *ḥis* 'blood, to bleed;' *ḥis-mis* 'blood/bleed-NMLZ'
f. 'saliva': *taaxckʷi* /taaxʷ-ckʷɨ̌/ 'spit-remains.of'
g. 'sweat': *ƛ̓upẏiiḥa-ckʷim* 'to.sweat-??'
h. 'snot': *ʃintmis* 'snot, nasal mucous'

The only meaning not (also) encoded as a basic verb is that for 'snot,' though note that it, too, ends in *-mis*, although there is no corresponding verb *ʃint* in the consulted source. Note that in Nuuchahnulth the root *ḥis*, which bears the semantic content of 'blood,' is ambiguous as to its lexical category and can function as both noun and verb, and that there exists an overtly nominalized version of this which singles out the referential read-

ing. The Nuuchahnulth source does not contain counterparts for the meanings ‘pus’ and ‘urine;’ verb-based terms for these meanings are found for instance in Chickasaw (*kalha-'* ‘have.pus.come.out-NMLZ’) and Sora (*'aŋ(ɲ)um-ən* ‘urinate-N.SFX’).

Percentages for terms like those in (10.) for all languages are in Appendix C (note that only plain, semantically inert derivation serving only to change the lexical category is counted here; thus Pipil *te:mal* ‘pus,’ which is derived from *te:ma-* ‘to fill,’ and similar terms are not counted, as are semianalyzable terms of all kinds).

The map in figure 38 shows the distribution and strength of the phenomenon visually. As the map shows, such terms are relatively frequent in North America, which can also be observed in the boxplot in figure 39. However, statistically, the areal differences are not quite significant under the Dryer-6 breakdown ($\chi^2 = 10.3366$, df = 5, $p = .06624$, Kruskal-Wallis rank sum test).

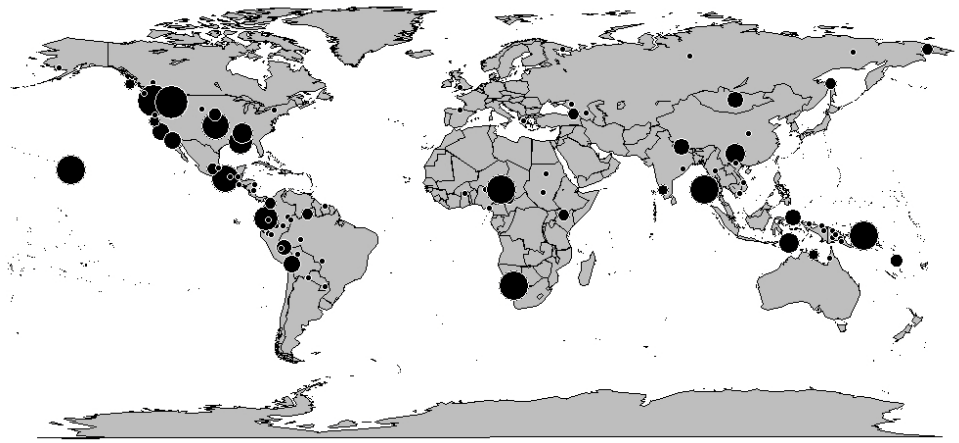


fig 38: the distribution of deverbal or N/V-ambiguous terms for aerosols and body liquids, core sample

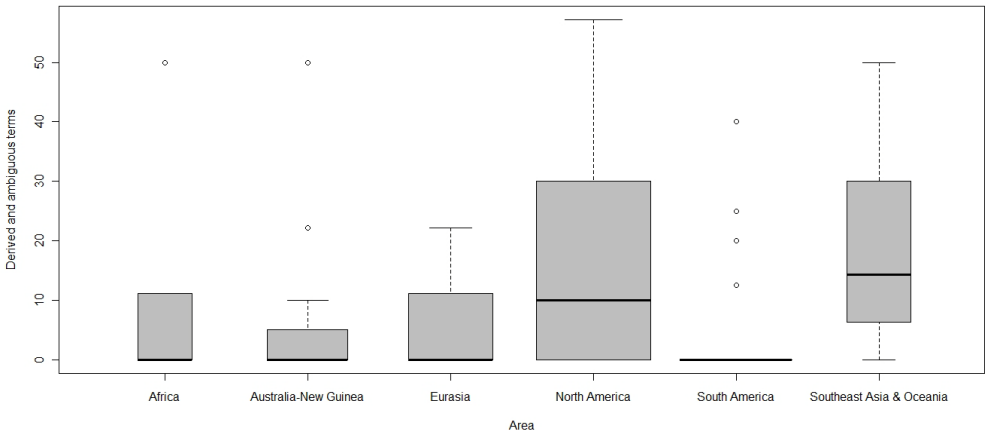


fig. 39: Percentage of deverbal or N/V-ambigous terms in different areas, using Dryer's (1992) breakdown

The picture is quite similar when only overtly marked deverbal terms are considered. The corresponding map is seen in figure 40.

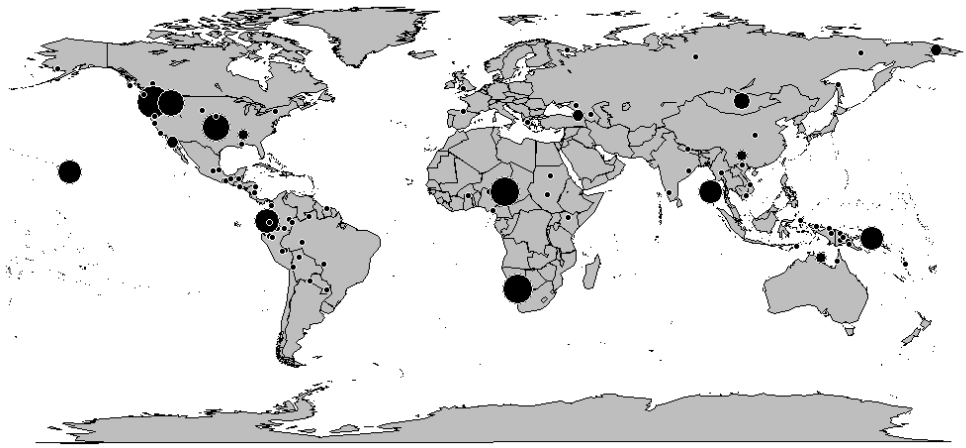


fig. 40: the distribution of deverbal terms for aerosols and body liquids, core sample

In spite of the fact that here North America stands out even more clearly when it comes to verb-based lexical categorization of the relevant meanings, as also seen in the boxplot in figure 41, areal biases are not significant statistically (χ^2 = 3.4827, df = 5, p = .626, Kruskal-Wallis rank sum test).

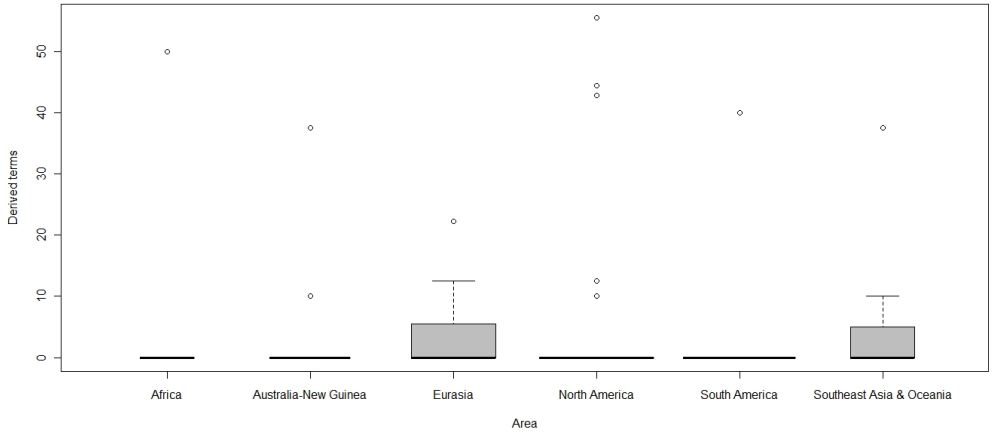


fig. 41: Percentage of deverbal terms in different areas, using Dryer's (1992) breakdown

Unsurprisingly, there is a strong correlation in both cases with the percentage of derived terms for the other meanings considered (Spearman's ρ = 3086653, p = .005968 when taking into account both derived and ambiguous terms and Spearman's ρ = .3461545 and p = .001907 when only taking into account derived terms).

5.6. A NOTE ON COLEXIFICATION, ANALYZABILITY, AND PHONOLOGY

Could it be possible that the phonological structure also is responsible for the degree of polysemy or at least for certain patterns of colexification, such as 'river' - 'water'? This is suggested both by the case study of Vanimo as well as the situation in Northwest Caucasian, where it has been noted that simplex lexical items have a rather broad denotational range to compensate for the limited number of roots the phonological system enforces. Thus it is a conceivable situation that a high degree of analyzability goes hand in hand with a high degree of simplex lexical items with broad reference, in other words, lexical items that colexify meanings that would be expressed by morphologically unrelated lexical items in other languages. As noted in chapter 3, next to more general issues having to do with the extraction of colexification, there is an effect of the type of the consulted source on the quantitative measure of colexification so that testing on the entire statistics sample is not feasible. However, it is possible to narrow down the sample even further by removing data from languages for which the source is of the kind that influences this percentage statistically. However, even with this measure taken, statistically no interaction of the percentage of colexification was found with any of the phonological features under scrutiny. In contrast, there is a correlation between the degree of analyzability and the degree of colexification (values for both are in Appendix B) for those languages where there is no statistical bias on the measured degree of colexification due to the nature of the consulted source. This analysis shows that, on average, languages with a high degree of colexifying lexical items also tend to have low degrees of analyzability, while languages

with a comparably low percentage score when it comes to colexification, typically have a more analyzable lexicon (p = .0018 by a Mixed Model design). Thus, rather than a upward trend in the degree of colexification that is correlated with a rise in the number of analyzable lexical items in the investigated vocabulary, there is an inverse relationship between colexification and analyzability: The more analyzability, the less colexification and vice versa. The correlation is plotted in figure 42.

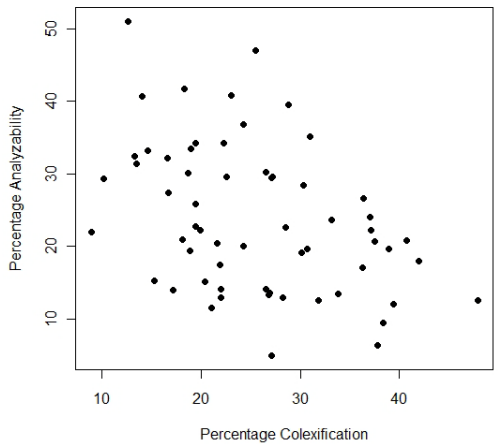


fig. 42: Correlation between the measured percentage of colexification and analyzability

This mirrors the basic observation from § 3.5.: the same semantic relationship may be expressed by colexification in some languages and by analyzable lexical items in others.

5.7. METAPHOR AND METONYMY

5.7.1. INTRODUCTION

Rather than looking at the quantitative aspect of lexical motivation, with which most of the discussion in this chapter has been concerned so far, this final section looks at the semantic side of things, in particular contrasting the degrees to which languages employ metaphor or metonymy as defined in chapter 3 as semantic relations. For quantitative evaluation, these differences are measured by the CONTIGUITY-SIMILARITY RATIO, which is calculated by dividing the relative percentage of lexical items motivated by similarity by the relative percentage of lexical items motivated by contiguity. Hence, a value of 1 indicates that the two values are in balance, a value smaller than one indicates that contiguity dominates (the smaller the value, the stronger this dominance is) and a value larger than one that similarity is the dominant semantic relation in a given language (again, the larger the value, the stronger the dominance). Values for this ratio are in Appendix B. The map in figure 43 plots the cross-linguistic differences in this area for the languages of the core sample.

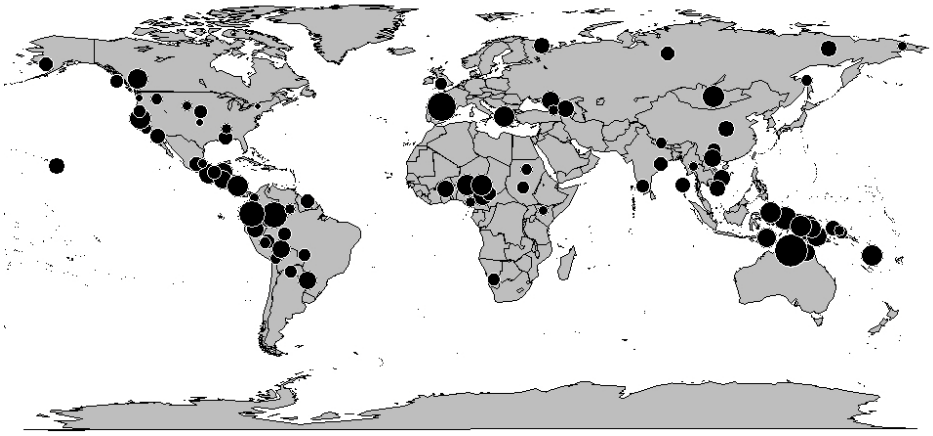
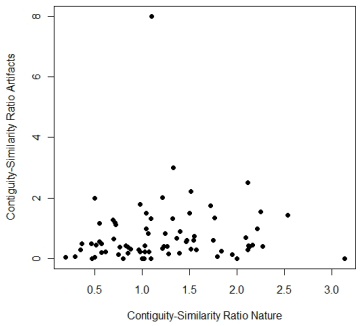


fig. 43: differential degree of metaphor- vs. metonymy in motivated terms, core sample

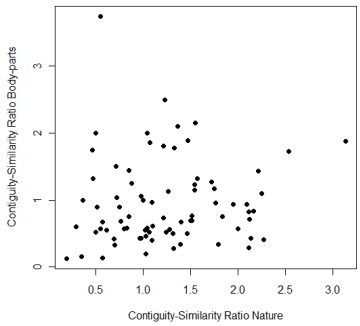
A question raised by Koch and Marzo (2007: 273) is: "Are there predominantly metaphorical languages?," in other words, whether there is any non-random signal in the distribution of the variable as seen in figure 43. The answer to this question is, as the following discussion will show, yes, there seem to be, but the much more interesting question to ask is, why?

5.7.2. CORRELATIONS BETWEEN THE PROFILE OF LANGUAGES IN DIFFERENT SEMANTIC DOMAINS

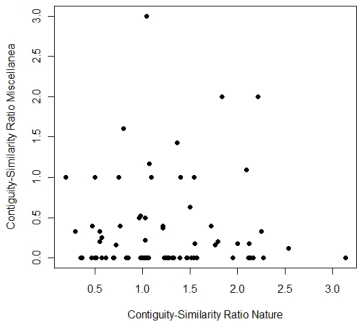
The plots in figure 44 visualize differences of metaphorical vs. metonymic semantic relations in the languages of the statistics sample across semantic domains (see Appendix B for data), and tests for each possible combination for correlations between the domains, with the Spearman's ρ being approximate due to ties and *p*-values corrected using Bonferroni corrections as implemented in R due to multiple testing.



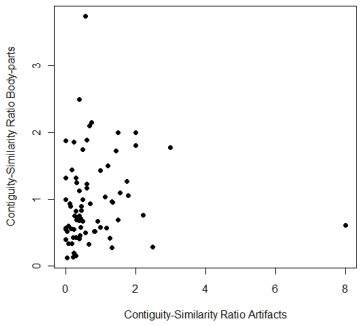
Nature vs. Artifacts: ρ ≈ .18, *p* ≈ .60



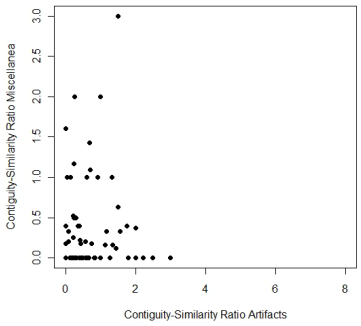
Nature vs. Bodyparts: ρ ≈ .14, p ≈ .94



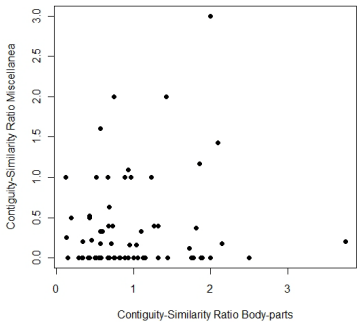
Nature vs. Miscellanea: ρ ≈ -.02, p = 1



Artifacts vs. Bodyparts: ρ ≈ .29, p ≈ .07



Artifacts vs. Miscellanea: ρ ≈ -.03, p = 1



Bodyparts vs. Miscellanea: ρ ≈ .09, p = 1

fig. 44: the contiguity-similarity ratio across semantic domains

What this analysis shows is that under all other possible pairings, the correlation is not significant, thus meaning that here the distribution is less clearly paired when comparing it with the results for the degree of analyzability reported in § 5.2.2.

However, next to asking about domains and the differential degree to which contiguity- and similarity-based denominations are found, it is also possible to ask whether there is any difference with respect to their subtypes as established in § 3.6.2.2. The boxplot in fig. 45[54] shows that the ratios of terms where the relation of functional similarity as opposed to perceptual similarity, is found to the highest degree is that of artifacts (values are in Appendix D).

[54] The extremely high value of 276.5 for the meaning 'bark' lies outside the plotted area.

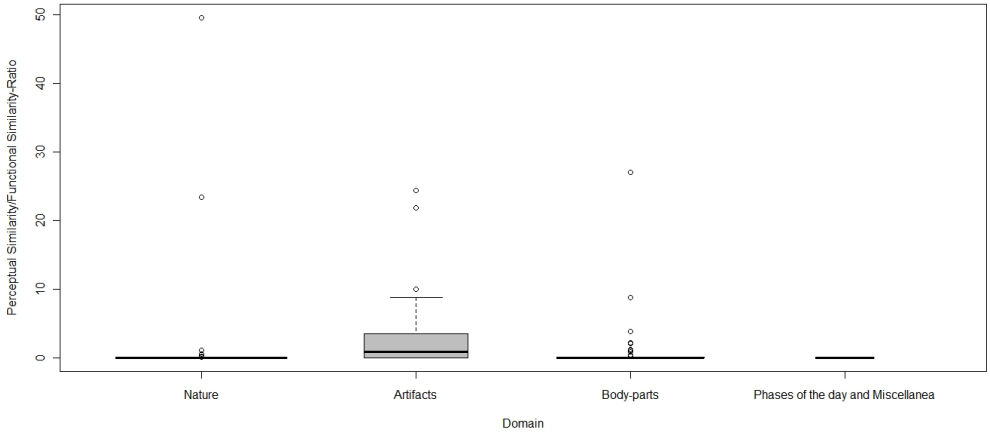


fig. 45: differences across semantic domains in functionally- vs. perceptually-based similarity

This is mostly due to the distribution of the relation of functional similarity within semantic subdomains. Compare, for instance, as examples of morphologically complex terms, Hausa *jurgi-n sama* 'boat/train-GEN sky' = 'airplane' as well as Kaluli *ho:n ko:su* 'water airplane' = 'power boat, boat introduced during colonial contact.' In the domain of tools, frequently colexification of meanings that have functionally similar referents are found, for instance Jarawara *yimawa* 'knife, machete' (cf. also Sko *tàng*, glossed as 'sickle, knife, machete, general term for blade of any kind' and note in this regard that the distinction between genuine polysemy and semantic generality is not at stake presently) and Sentani *o'bi* 'ladder, stairs.' Another frequent pattern is colexification of 'house' and 'nest' (in spite of the equally if not more common pattern for 'nest' to be named by a morphologically complex term 'bird house,' see Appendix E, 41), and abstract extension of meanings such as that of 'street' or 'way' to 'method' or 'manner' (see Appendix E, 92).

A similar result, with artifact-terms standing out, is obtained when one does not look at the difference between the two different types of similarity-based relations, but instead compares for each concept whether contiguity-based or similarity-based conceptualizations, as measured by the contiguity-similarity ratio, abound. As the plot in figure 46[55] shows, it is again the domain of artifacts in which particularly low values for the contiguity-similarity ratio, that is, prevalence of contiguity as the semantic relation is found.

[55] Again, the meaning 'bark' is with a value of 111 outside the plotted area.

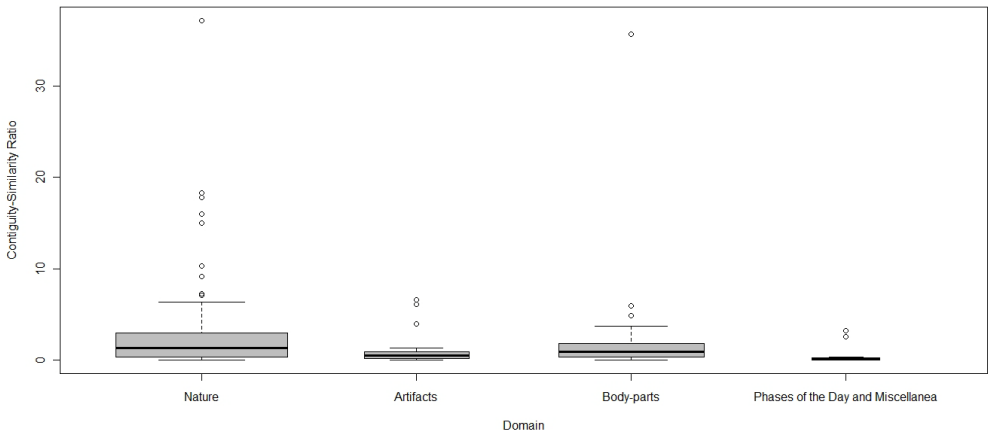


fig. 46: differences across semantic domains in prevalence of contiguity and similarity-driven conceptualizations

However, the miscellanea-domain shows a similar behavior, which is in all likelihood due to contiguity-based associations for meanings such as 'day' and 'night.' For instance, 'day' is, by contiguity, frequently colexified with 'sun,' and 'night' with 'dark.' Moreover, contiguity-based terms for 'noon' such as Kildin Saami *piejjv-kēssk* 'day-middle' abound (see Appendix E, 151, 153, and 154 for fuller discussion).

5.7.3. INFLUENCES OF STRUCTURAL FACTORS?

Analogously to the data on the degree of analyzability and the type of analyzable lexical item, preliminary tests were carried out on the basis of the data in the World Atlas of Languages Structures to elucidate possible interactions between structural features and the dominance of either contiguity or similarity as the semantic relation underlying analyzable items and colexification. Significant *p*-values (all by Kruskal-Wallis Rank sum tests, and again, given the exploratory nature of the tests, uncorrected for multiple hypothesis testing) were obtained for the following features:

(i) Voicing in Plosives and Fricatives (Maddieson 2005g):
$\chi^2 = 10.1558$, df = 3, $p = .01729$
(ii) Uvular Consonants (Maddieson 2005f):
$\chi^2 = 7.2236$, df = 2, $p = .02700$
(iii) Lateral Consonants (Maddieson 2005c):
$\chi^2 = 11.9302$, df = 4, $p = .01788$
(iv) Politeness Distinctions in Pronouns (Helmbrecht 2005):
$\chi^2 = 3.7231$, df = 1, $p = 0.05367$
(v) The past tense (Dahl and Velupillai 2005):
$\chi^2 = 8.6491$, df = 2, $p = 0.01324$

(vi) Order of relative clause and noun (Dryer 2005e):
χ^2= 8.1259, df = 3, p = 0.04348
(vii) Position of Interrogative Phrases in Content Questions (Dryer 2005h): χ^2= 5.4938, df = 2, p = .06413
(viii) Relationship between the Order of Object and Verb and the Order of Relative Clause and Noun (Dryer 2005k):
χ^2= 9.2295, df = 3, p = 0.02639
(ix) Alignment of case marking of pronouns (Comrie 2005):
χ^2= 10.5261, df = 5, p = 0.06163
(x) Zero Copula for predicate nominals (Stassen 2005b)
χ^2= 3.7386, df = 1, p = .05317
(xi) Tea (Dahl 2005): χ^2 = 5.7909, df = 2, p = 0.05527

Of these, all but three features remained significant under a Mixed Model design controlling for areal effects. The remaining eight features are:

(i) Voicing in Plosives and Fricatives (Maddieson 2005g) : p = .0209
(ii) Uvular Consonants (Maddieson 2005f): p = .0172
(iii) Lateral Consonants (Maddieson 2005c): p = .0263
(iv) Politeness Distinctions in Pronouns (Helmbrecht 2005): p = .0378
(v) The past tense (Dahl and Velupillai 2005): p = .0029
(vi) Order of relative clause and noun (Dryer 2005e): p = .021
(vii) Relationship between the Order of Object and Verb and the Order of Relative Clause and Noun (Dryer 2005k): p = .0067
(viii) Alignment of case marking on pronouns (Comrie 2005): p = .0113

Cross-validating the results on the basis of the validation sample was not possible for features (i), (ii), (iii), (v), and (viii).[56]

There was a replicable difference between languages with no and a binary politeness distinction in pronouns (estimate of the validation sample 0.46 as opposed to 0.4503 ± 0.1862 in the original sample), which is plotted in fig. 47.

[56] The available estimates for the sake of completeness are: (i): -.1650 vs. .346 ± .1236, .08 vs. .3627 ± .2141 and .06 vs. .6293 ± .2141; (ii): -.2533 vs. -.40640 ± .14934; (iii): .5333 vs. .27476 ± .14366 and .185 vs. -.14190 ± .22714; (v): .02 vs. -.678 ± .1442; (viii): .0660 vs. .395 ± .1225.

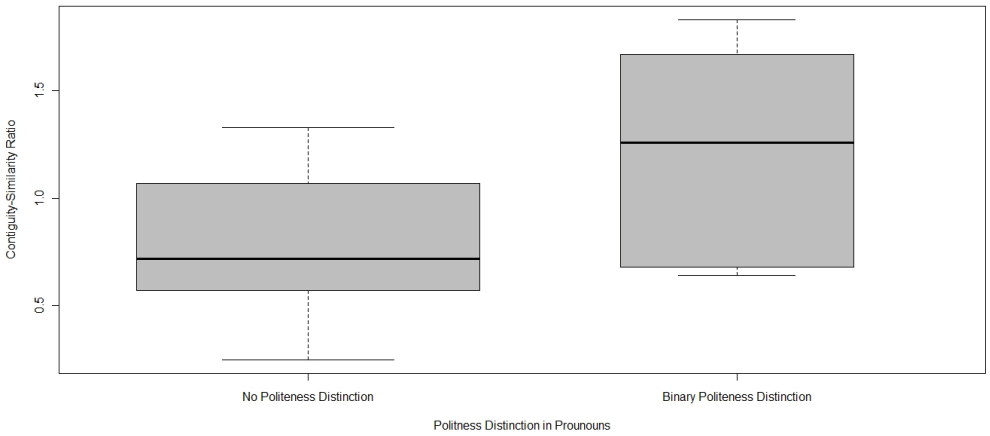


fig. 47: differences in the contiguity-similarity ratio depending on politeness distinctions in pronouns

Moreover, there are two replicable correlations which have to do with the order of relative clause and noun. As the associated plots in figures 48 and 49 show among other information, metaphor-based associations are more common in languages in which relative clauses precede the noun.

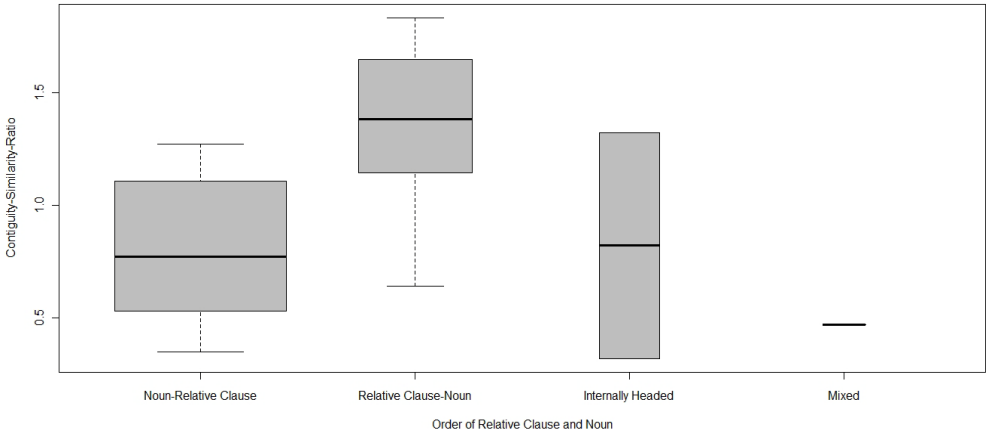


fig. 48: differences in the contiguity-similarity ratio depending on the order of relative clause and noun

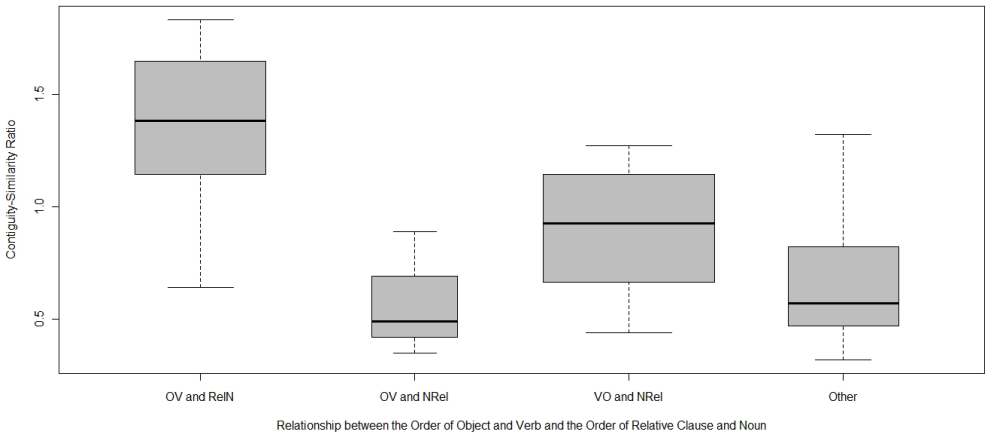


fig. 49: differences in the contiguity-similarity ratio depending on the order of object and verb and the order of relative clause and noun

Generally, these correlations are difficult to make sense of, and hence they are for the time being simply mentioned without an attempt at an explanation. It needs to be borne in mind that, as already noted in the discussion above, the overlap between the WALS samples and the present sample is at times rather small, and hence so is the empirical datapool from which generalization may be drawn, to the effect that the behavior of few individual languages can lead to the emergence of statistical significance. Conversely, also because of these facts, some genuine interaction in fact may exist for features for which none has been diagnosed. At any rate, with politeness distinctions in pronouns and the order of relative clause and noun as the only candidates, the influence of structural features coded in WALS on semantic relations underlying motivated items in the lexicon appears to be small.

5.7.4. NO STRONG AREAL EFFECTS ON THE RELATIVE DEGREE OF METAPHOR AND METONYMY

Areal effects are not very pronounced either, and where they exist, there are relatively straightforward explanations. Figure 50 plots the results of the relative degree of metaphorical expressions using Dryer's 6-way breakdown of the world. This is for the time being simply for the purpose of illustration; the difference is not significant statistically ($\chi^2 = 5.2236$, df = 5, $p = .3892$, Kruskal-Wallis rank sum test).

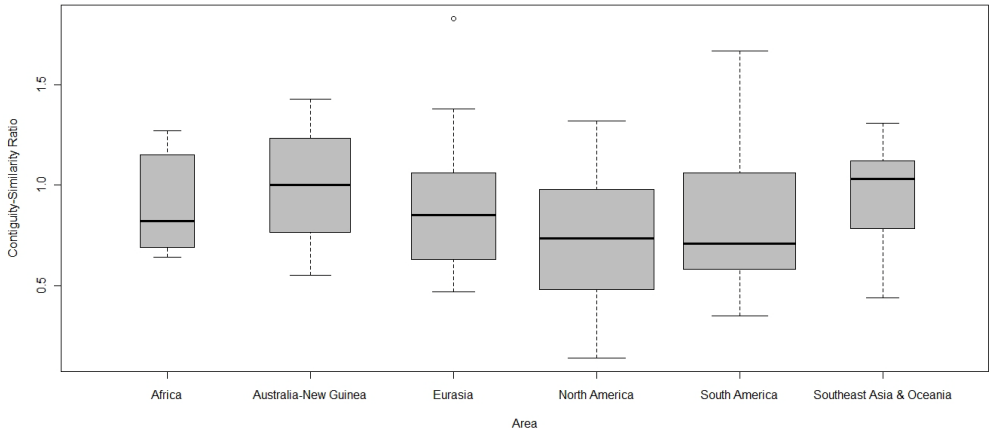


fig. 50: Areal breakdown of the relative degree of metaphor-driven semantic relations, using Dryer's (1992) breakdown.

The lowest degree of metaphor-based conceptualization is found in the Americas, where contiguity as a semantic mechanism in colexification and analyzable lexical items prevails, although it is not significantly more dominant here than elsewhere. There was also no evidence for areal differences under the other two standard breakdowns used in the present study (Nichols-11: $\chi^2 = 11.9601$, df = 10, $p = .2877$; Nichols-3: $\chi^2 = 3.8834$, df = 2, $p = .1435$, both by Kruskal-Wallis rank sum tests).

Testing for individual semantic domains yields almost always negative results under all testing conditions, with the exception of artifacts, which have a significantly different relative degree of metaphor and metonymy at $p = .02277$ ($\chi^2 = 7.5644$, df = 2, Kruskal-Wallis rank sum test) under the Nichols-3 breakdown, plotted in figure 51.

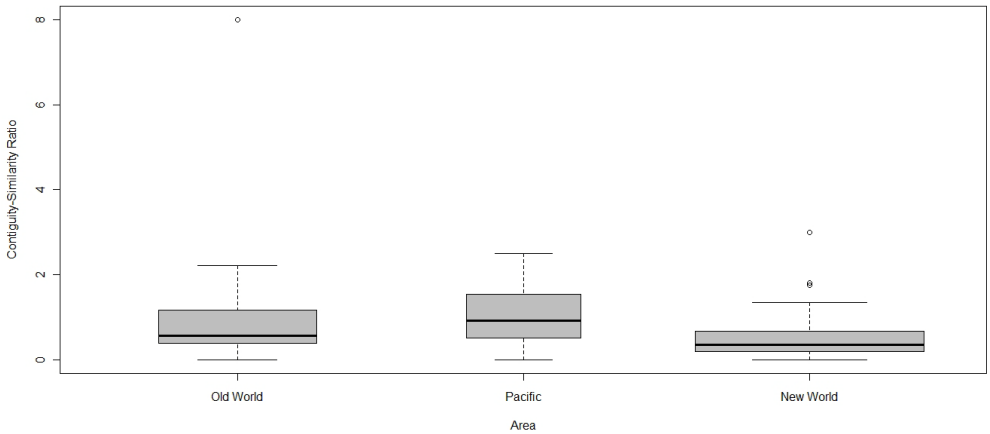


fig. 51: Areal breakdown of the relative degree of metaphor-driven semantic relations in artifacts, using Nichols's (1992: 27) breakdown

This plot essentially replicates the one when all semantic domains are considered: high degree of metaphorical relations in the Pacific area, with depressed ratios in the New World. This is most likely due to two factors: first, the languages of the Americas have a high ratio of motivated, in particular analyzable terms for artifacts, and these tend to be named with reference to their function, which is by definition a relationship of contiguity, not one of similarity. Perhaps more importantly, as will be seen in the following section, there is an overall correlation between the preference for terms of the derived rather than of the lexical kind to be driven by contiguity, and it is again in the Americas where languages of this type cluster.

This negative outcome should not be too surprising, if one bears in mind the ultimate cause of areal effects: the need of bi- or multilingual speakers to increase intertranslatability between the languages they speak (e.g. Gumperz and Wilson 1971), and the need to express the same thought in two different languages (Sasse 1985). As noted by both Gumperz and Wilson and Sasse, this single need underlies contact phenomena in morphology and syntax, but are equally responsible for convergence in semantics and lexicon. Since relative degree of metaphor is a highly abstract measure, it seems unlikely to be influenced by areal factors as it is not directly manipulateable by speakers. Rather, contact effects are clearly recognizable in the denominations of individual meanings and their semantic structure (see § 6.4.3).

5.7.5. METAPHOR AND METONYMY AND PREFERRED TYPE OF ANALYZABLE LEXICAL ITEM

If the degree of metaphor and contiguity does not appear to be decisively influenced by grammatical factors nor for the most part by areal factors, is their distribution completely random? In fact, there appears to be a structural factor that triggers the languages' behavior in this regard.

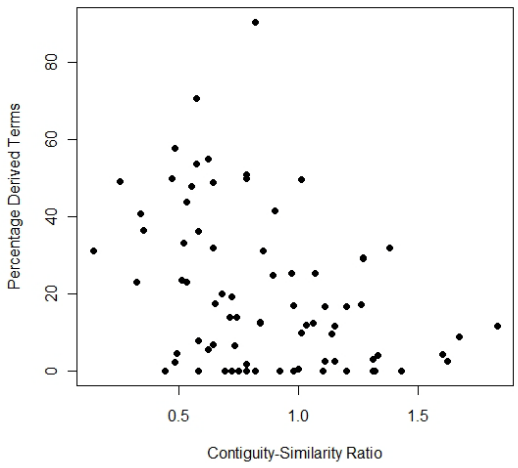


fig. 52: Correlation between the contiguity-similarity ratio and the percentage of derived terms of all analyzable terms

There is a non-trivial and significant effect of the relative frequency of complex lexemes of the derived and of the lexical type as defined in § 3.6.1. and the fundamental semantic relation -similarity or contiguity- that dominates the lexicon (*p*-value associated with predictor = .0004 by a Mixed Model design controlling for areal affects, estimate = -.0066). An illustrating plot is in figure 52.

Why is that? As mentioned in § 4.4.2., the Central Yup'ik postbase *-yak* 'thing similar to' is by its semantics prone to create complex lexical items where the referent of the complex terms stands in a relation of similarity to that of the derivation base, so it is logically perfectly possible to have similarity-based terms derivatives. Some examples from another language, Muna, which is one of the rare languages that have several terms of this kind distributed over all semantic domains, are in (11.):

(11.) Similarity-based derivatives in Muna
a. *ka-mbea* 'ABSTR-shine' = 'flower'
b. *ka-ofe ~ ka-ufe* 'ABSTR-squeeze.rice.in.round.shape' = 'nest'
c. *kara-kara* 'yard.fence-RED' = 'rib'

However, derived terms in most languages are not metaphorical in nature, but have a metonymic basis (see also Anderson 2011b: 285). This lies in the very nature of the process, more precisely, in the semantics of derivational morphemes found in many languages that often serve to derive names for instruments or locations from the derivation base (see Bauer 2002 for a cross-linguistic survey of the semantics of derivational morphemes, which includes a number of more unusual meanings, but none that is susceptible to establish a relation of similarity with the meaning of the derivation base in particular). Furthermore, derivatives typically do not allow for contiguity anchoring, leading to a "cognitive leap" that appears to be dispreferred cross-linguistically. Nuuchahnulth derivatives may serve as examples for the overwhelmingly contiguity-based derivatives in the world's languages:

(12.) Contiguity-based derivatives in Nuuchahnulth
a. *maamaati* /maa-mat-ǐʻp/ 'RED-fly-THING...ED' = 'bird'
b. *hiɬ-waḥsuɬ* 'LOC-go.out.from' = 'estuary'
c. *ƛupkẏak* /ƛupk-ẏak^{w}/ 'untie-INSTR' = 'key'

And even in Muna, the locative nominalizer *ka-* occurs typically in contiguity-establishing function, such as in *ka-bhawo* 'mountain' (*bhawo*, 'high').

However, the correlations are found for semantic relations in the lexicon as a whole, that is including those in morphologically complex lexemes as well as those in colexification. Since the above observations pertain exclusively to analyzable items, it is necessary for the present purposes to distinguish between semantic relations in analyzable items and those in colexification and to assess their behavior separately. When this is done the picture becomes much clearer. Then, there is a highly significant correlation at p = .0001 between the dominant type of complex lexical item (derived vs. lexical) and the predilection for similarity-driven as opposed to contiguity-driven semantic relations in analyzable terms (data are in Appendix B) under the same Mixed Model design controlling for area, with the value for the contiguity-similarity ratio logarithmically transformed. Figure 53 illustrates the correlation.

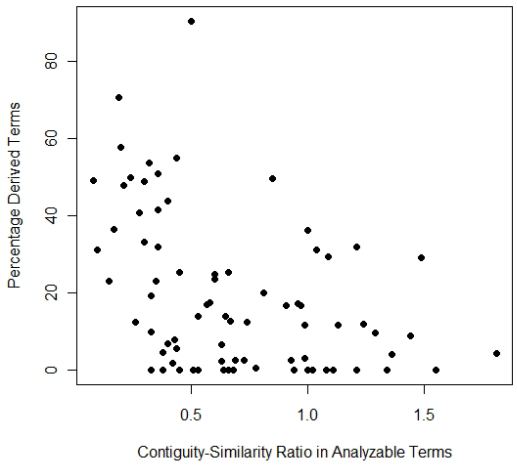


fig. 53: Correlation between contiguity-similarity ratio and the percentage of analyzable terms of the derived type

But what about colexification? It would be a spectacular finding if the same preference as in analyzable lexical items would extend to colexification as well. However, this is not so. There is no effect of the dominant type of complex lexical items on the semantic relation in colexification (p = .8858) under a Mixed Model design with the contiguity-similarity ratio in colexifying terms (values are again in Appendix B) logarithmically transformed, showing that the overall correlation between the variables is entirely due to the semantic relations in analyzable terms, with the relations due to colexification in fact confounding the picture.

It is possible to align this finding with the previous discussion, since both lack of verbal person marking and lack of an elaborate derivational apparatus are characteristic of a language type known as "isolating" in traditional morphological typology. Such languages, because of their restricted bound morphology, will to a great degree make use of lexical rather than derivational resources to coin their morphologically complex expres-

sions. In this sense, the observed patterns point to the TENDENCY OF ISOLATING LANGUAGES TO MAKE USE OF METAPHOR AS A LEXEME-INTERNAL SEMANTIC RELATION IN ANALYZABLE TERMS TO A GREATER DEGREE THAN NON-ISOLATING ONES.

The differences can be exemplified by contrasting data from Austro-Asiatic languages. This family consists of two major branches, the Munda languages spoken on the Indian subcontinent, and the Mon-Khmer languages, most of which are spoken in Southeast Asia. The split between these two primary branches is deep, and consequently, Munda and Mon-Khmer languages are also quite different typologically. Munda languages have rich verbal morphology; in contrast, Mon-Khmer languages participate in the Southeast Asian Sprachbund and exhibit its typical features: they are tonal and are largely isolating, thus for instance not featuring person agreement on the verb. These typological differences are mirrored in the lexicon, and they can be shown by contrasting data from Sora (Munda) and Sedang (Mon-Khmer). In Sora, about 25 per cent of analyzable terms in the data are of the derived type. Examples include:

(13.) a. *'ge:mən* /ge:m-ən/ 'to.light-N.SFX' = 'flame'
b. *gərob'go:b-* /g<ər>ob-gob-/ '<INSTR>sit-RED-'= 'seat'
c. *meme:-n* 'suck-N.SFX' = 'breast'
d. *gag'garən ~ gal'galən /gag-gar-ən/* 'RED-pierce/bore.a.hole-N.SFX' = 'scar'

(Of course, Sora also features complex expressions that are not based on verbs, such as *'kuru:-tam-ən* 'body.hair-mouth-N.SFX' for 'beard'). In the lexicon in general as in the examples in (13.), contiguity-driven conceptualizations outnumber metaphor-driven ones, as indicated by the contiguity-similarity ratio of .79 (.4 in analyzable terms only).

In line with the observations made above, there is a correlation between dominance of complex expressions of the lexical type and metaphor as the conceptual mechanism underlying lexical motivation. In the Mon-Khmer language Sedang, one encounters roughly the reverse situation. The percentage of derived terms is, with 11.9 per cent of all analyzable terms, only about half of that encountered in Sora. In absolute figures, this amounts to a number of only two deverbal terms in the data available for Sedang, formed using the nominalizing infix <ơn>, for instance *kơnep* 'scissors' (*kep*, 'to cut hair'). Examples of analyzable terms of the lexical type in Sedang are in (14.).

(14.) a. *kia hia* 'ghost light.weight' = 'clouds, air, smoke'
b. *kơtôu ma* 'bark/rind/shell eye' = 'eyelid'
c. *tróang mơhéam* 'road blood' = 'blood vessel, vein, artery'
d. *tea ma* 'water/liquid eye' = 'tear'

The chosen examples are roughly representative of the relative degree of contiguity and similarity as underlying processes: while there clearly are complex terms that are contiguity-driven, such as (14d.), similarity-based complex lexemes outnumber them in the vocabulary segment under investigation, as indicated by the contiguity-similarity ratio of 1.03 (1.24 in analyzable terms only).

5.7.6. A FURTHER POSSIBLE FACTOR

There is a strand of recent research in Social Psychology that may turn out to open up extremely interesting prospects for a better understanding of preferences between languages for the prevalence of semantic associations they favor. Cognitive Psychologists distinguish two types of reasoning, one is based on intuitions on the basis of gathered experience and is associative in nature, the other is categorical, logical and operates by the application of rules (Sloman 1996). The former system is based on relations of spatiotemporal contiguity and similarity, the latter on categorical and taxonomic relations. Importantly, although both systems are probably available to all humans, there are marked cross-cultural differences in the prevalence of each of the systems in reasoning. In particular, the former, associative system is dominant in Asia, while in languages of Western cultures, the taxonomically oriented system is employed with greater frequency (Norenzayan et al. 2002, Nisbett 2003, see Norenzayan et al. 2007: 577-586 for review). For instance, Masuda and Nisbett (2001) show that Japanese subjects remember more of the background of an artificial underwater scene they were shown, and started descriptions of the scene by introducing the background, Westerners were more likely to separate a particularly salient target object - a "focal fish," which is bigger and more colorful than other elements - of the scene. Ji et al. (2004) show that prevalence of one of the two systems has effects in linguistic tasks specifically: in a triad categorization task, American subjects were more likely to group sets of words together on the basis of category structure (for instance, grouping 'monkey' together with 'panda'), while Chinese-speaking subjects were more likely to group referents together on the basis of them sharing the same frame (e.g. 'monkey' and 'bananas'). The Chinese subjects had some degree of proficiency in English, and were tested using both English and Chinese; the effects remained noticeable regardless of this difference.

Thus, given the areal distribution of the dominance of the systems, it may be the case that in languages of Western cultures, motivated terms, in particular neologisms, may be characterized by reflecting taxonomic structures, as e.g. in endocentric compounds, while denominations in Asian languages could be expected to be of an associative, contiguity and/or similarity-based (as e.g. in exocentric compounds). In the areal breakdown in figure 50, one can observe that there is a higher number of metaphor-driven lexical associations than in languages of Eurasia and Europe. However, it is not entirely clear whether the distinction of contiguity vs. similarity as presently defined is in fact the adequate measure to bring to light such putative influences in language, since the associative system operates both on the basis of spatiotemporal contiguity and family resemblances (metaphor), and it might be more profitable indeed to approach the question distinguishing between e.g. endocentric and exocentric compounds.

Moreover, the question of whether there are indeed cross-linguistic effects of the two types of reasoning on lexical structure unfortunately cannot at this point of time be elucidated in more detail because "[l]ittle is known about the operation of these two systems of reasoning across diverse cultural groups" (Norenzayan et al. 2002: 654), in spite of some evidence that, rather than a difference between Western and East Asian societies, on contrasting which research has focussed so far, the difference really is between the industrialized West and the rest of the world as well as differences based on mode of subsis-

tence (Henrich et al. 2010). However, these differences in cognitive styles (to take up a term by Hymes 1961) and their effects on linguistic tasks demonstrated by Ji et al. (2004), which are in turn likely based on different patterns of social, political, and personal organization (Nisbett et al. 2001), suggest that it is possible that there are cultural effects on the structuring of the lexicon that would provide evidence against the claim uttered for instance by Alinei (2001) that languages randomly pick features of referents in naming them and that goes beyond the trivial sense of contingent aspects of material culture, as when, say, a language colexifies 'thorn' and 'needle' because thorns are used as needles.

5.7.7. SUMMARY

Given that influences of cognitive reasoning cannot be systematically checked at the present state of knowledge, the overall conclusion for the time being thus is that the DOMINANT WORD-FORMATION DEVICE INFLUENCES WHETHER THE LANGUAGE WILL FAVOR CONTIGUITY- OR SIMILARITY-BASED DENOMINATIONS IN MORPHOOGICALLY COMPLEX LEXICAL ITEMS. This is a non-trivial finding, since, to reiterate, there is no a priori reason that compounds must be metaphorical and derivatives must be metonymic semantically. It is also a highly interesting finding because, put in other words, one can observe here that languages, depending on the nature of aspects of their grammar (i.e. word-formation), carve up the essentially same or near-same reality, as represented by the meanings on the list which are presently studied, in quite different ways. At any rate, it would be highly interesting to expand the findings empirically in concrete fieldwork to ascertain the soundness of the semantic side of the analysis. As pointed out by Aikhenvald (2007: 9), "compounding is widespread in isolating languages, while derivation is a property of languages of other types; this follows from the tendency to have a one-to-one correspondence between a morpheme and a word in isolating languages." It is therefore no coincidence that high rates of contiguity-based semantic relations at the expense of similarity-driven ones are dominant in the Americas, because here derived-type languages concentrate (though note that the correlation is not due to this fact alone, since area is controlled for). The typology can now be enhanced and finalized in table 26 by adding a lexico-semantic correlate to the lexical and derived types: that of predilections for similarity-based and contiguity-based semantic relations in morphologically complex lexical items respectively.

	High degree of Analyzable Terms	Low Degree of Analyzable Terms
Lexical Dominating, Derived Subsidiary	• Low complexity in verbal person marking, fixed word order	• Low complexity in verbal person marking, fixed word order
	•Simple phonology, short roots	•Complex phonology, long roots
	•Dominance of similarity as a semantic relation in analyzable terms	•Dominance of similarity as a semantic relation in analyzable terms
	•Tentatively: favors neologisms	•Tentatively: favors borrowing
Derived Dominating, Lexical Subsidiary	• High complexity in verbal person marking, free word order	• High complexity in verbal person marking, free word order
	•Simple phonology, short roots	•Complex phonology, long roots
	•Dominance of contiguity as a semantic relation in analyzable terms	•Dominance of contiguity as a semantic relation in analyzable terms
	•Tentatively: favors neologisms	•Tentatively: favors borrowing

table 26: final cross-classification of language types summarizing the established correlations

With this table, the quantitative evaluation comes to an end. It is summarized in textual form in the final section that is to follow.

5.8. CHAPTER SUMMARY

This chapter presented a quantitative evaluation of the variables surveyed in this work, and tried to establish correlations with language-internal structural as well as some social and cognitive factors and to provide, where they are found, an explanation for the observations. It turned out that most of the obviously relevant factors that interact with the degree of analyzability of the nominal lexicon is structural rather than areal-typological (borrowing etc.), sociolinguistic (L2 learners) or cultural (word taboo, mode of subsistence). More precisely, the structural factors involved are mostly phonological: the simpler the syllable structure, the smaller the consonant inventory, the shorter the monomorphemic native lexical morpheme, the more analyzable terms the sample languages have in their nominal lexicon. Another relevant phonological factor is suprasegmental: tone, such that tonal languages are likely to be characterized by a higher degree of analyzability in the lexicon that non-tonal ones. Each of the factors alone was found to be significant, but due to cross-linguistic dependencies between themselves, it is not entirely clear whether any of them has more weight than another or whether they "team up" and jointly exert influence on the structure of the lexicon. At any rate, when all factors are conflated into a single index of complexity, the correlation with analyzability in the lexicon that is observed is very strong. Thus, taken together, the identified factors together jointly account for the behavior of the sampled languages, and by means of them, it is

possible to extrapolate from the sample on the entire population of languages presently spoken and to make some predictions about their behavior. Intra-family comparison revealed that often the same dependencies that are observed in inter-language comparison hold, that is, genealogically related languages are subject to the same trend. As a candidate for a functional motivation for the correlations, homonymy avoidance was discussed, though there are difficulties in demonstrating how precisely this putative principle operates, and it may be that a less strong, but more reliable, case can be made for a weak functional principle that balances off between phonological and lexical complexity, which are poles of a continuum on which languages place themselves somewhere along the axes.

As for predilections for either metaphor- or contiguity-based conceptualization of the investigated meanings, the main relevant factor turned out to be differences in the favored word-formation device. Languages with many derived terms favor, by the nature of the process, contiguity-driven relations in analyzable terms, while languages with more analyzable terms of the lexical type tend to have more metaphor-based denominations. There are other structural factors for which a statistically significant influence can be observed, but the functional connection of them to this variable are unclear. In addition, several smaller sections and excursuses, some of which have to be seen as preliminary investigations of an at times speculative nature, were devoted to topics such as analyzability in reconstructed proto-languages (with particular reference to Indo-European), to differential degrees of borrowing in languages of the Americas, and to differences between languages in the lexicon in noun- as opposed to verb-based orientation.