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

Urban, M.

Chapter 4

The Formation of Complex Lexemes: Types and Strategies

4.1. INTRODUCTION

The aim of this section is to give an overview of the impressive variety of different types of morphologically complex nominals that can be found in the world's languages, and thus to give a "feel" for the data to be analyzed in a more quantitative fashion in the subsequent chapter. At the same time, a subsidiary goal of this section is to introduce and explain the basic statistical techniques of data analysis that will be used in doing so. Since this study is both interested in areal as well as universal typological trends and correlations, what is required is (i) a method to assess areal patterns in a meaningful way and (ii) statistical tools for determining valid typological trends, that is, trends which are precisely not due to areal tendencies in the data. These techniques will play a major role in the next chapter which is concerned with a quantitative analysis of the gathered data. This section, on the other hand, introduces these techniques embedded in a more traditional typological discussion in order to not overwhelm the reader with a lengthy technical statistical discussion at one point in the thesis, but rather to introduce relevant methods of analysis in a piecemeal fashion to gradually familiarize him/her with them.

With respect to the main goal of this section, it must be emphasized that its intention is not to offer a full-fledged elaborate typology of complex nominals on the basis of their internal structure. Such a typology, with particular reference to polysynthetic languages, is developed in Mattissen (2003, 2004). Mattissen's basic distinction is between an affixal and a compositional type of morphologically complex nouns, in which the former concatenate "non-root bound morphemes" around a lexical root, while in the latter type more than one lexical root can be combined in an ad-hoc fashion. Languages of the latter structure may be subdivided into a scope-ordered and templatic type. Her typology also offers a survey of different functional and semantic categories that may be incorporated into the nominal form; however, one important difference to the present study is

that she is less concerned with the institutionalized lexical level, but rather with complexity in noun forms that may arise spontaneously in discourse (Mattissen 2003: 254-255). Furthermore, Bauer (2009) offers a typology of compounds, and Bauer (2002) investigates the semantics of derivation from a cross-linguistic perspective.

While situated in the simple present framework for the typologization of formal relations in analyzable lexical items, this section aims instead to add flesh to this rough typologization by pointing out language types that rely on one kind of complex nominal more than others and by providing examples from particular sampled languages and the language-particular word-formation devices they employ. Another aim in doing so is to highlight some of the more unusual devices for the formation of nominals that are typically not discussed in textbooks and other standard accounts of word-formation, such as nominal classifiers as well as clausal nominals as found predominantly but not exclusively in languages of North America.

4.2. A FIRST TYPOLOGY

The two very rough types of analyzable lexemes as defined in chapter 3 are those of the lexical and those of the derived type, and this classification will be the starting point for the presentation of the results. There is wide variation in the degree to which the languages of the world rely on one or the other basic type of morphologically complex lexemes. Central Yup'ik, for one, relies virtually exclusively on the derived type. On the other hand, there is a quite large number of languages in which complex lexemes are exclusively of the lexical type, as well as many languages which exhibit a mixed behavior, where both morphologically complex lexical items of the derived and lexical type are found to varying degrees. As can be seen from the histogram in figure 1, the lexical type is much more frequent on a global scale than the purely deriving type (though, as will seen below, there clearly are areal factors involved), and for those languages that fall in between the extremes, those in which items of the lexical type are more common than those of the derived type are greater in number than those with the reverse situation.

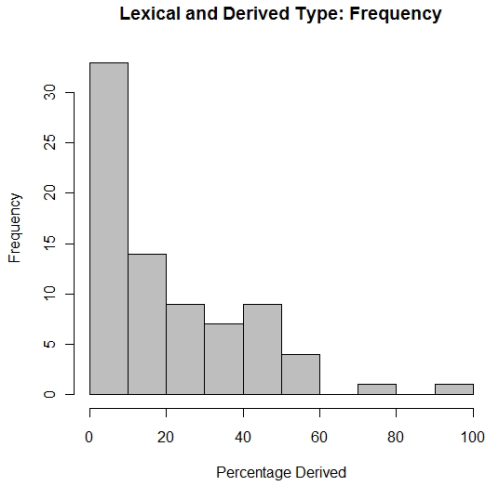

fig. 1: histogram showing the frequencies of languages with various degrees of complex items of the derived type globally

As a first approach to typologizing languages, the relative degree of derived terms in the languages of the core sample is divided into four quartiles, with languages of the first quartile (those with a preponderance for analyzable lexemes of the lexical type) constituting the first typological group, and those in the fourth quartile (those with a preponderance for analyzable lexemes of the derived type) forming the second. These two groupings define extreme types, and obviously there are many languages (those in the second and third quartile) that fall somewhere in between, that is, that have varying degrees of a mixture of analyzable items of the lexical and of the derived type. The third type of motivated lexeme which at least in some cases involves analyzability is the alternating type as defined in § 3.6.1. However, this device, compared with the lexical and the derived types, is so rare cross-linguistically that it is practically negligible for typologization (though there will be some discussion of gender alternation, the most common subtype of the alternating type in the following). This entails that the counts for the derived type and for the lexical type are practically mirror images of each other: where the derived type dominates, the lexical type is not pronounced, and vice versa.

This classification may seem a little ad hoc, but as will become clear later in the discussion, there are some noticeable typological correlates and commonalities between the languages of the extreme types.[1] Table 1 summarizes this rough typology, and provides an overview of the discussion concerning individual types that cluster at either of

[1] In general though, the ensuing discussion will be somewhat liberal given that what one is dealing with is really a continuum. Thus, languages whose percentages of derived terms are at the very top or bottom of a certain quartile are sometimes discussed along with the respective higher or lower group.

the extremes of the distribution (for percentages in the individual languages see Appendix B).

	Lexicon	Languages (in decreasing order of dominance of the derived type)	Typological profiles and reference to further discussion
Quartile I:	Derivational type dominant, lexical type subsidiary	Central Yup'ik, Burarra, Bora, Nunggubuyu, Xicotepec de Juárez Totonac, Ineseño Chumash, Nez Perce, Carib, Chukchi, Cashinahua, Chayahuita, Oneida, Buin, Rotokas, Aguaruna, Mali, Upper Chehalis, Cubeo, Aymara, Greek, Yanomámi, Efik, Khoekhoe, Khalkha	Languages with nominal classification (§ 4.4.1) Affixal type of polysynthesis (§ 4.4.2)
Quartile II + III	None of the types dominant more derived terms more lexical terms	Great Andamanese, Nuuchahnulth, Hausa, Sahu, Sora, Pipil, Arabela, Kiliwa, Laz, Kiowa, Chickasaw, Wintu, Hawaiian, Welsh, Haida, Wichí, Carrier, Kolyma Yukaghir, Kanuri, San Mateo del Mar Huave, Guaraní, Pawnee, Jarawara, Biloxi, Cavineña, Santiago Mexquititlan Otomí, Sedang, Basque, Itzaj, Bezhta, Tetun, Imbabura Quechua, Piro, Bororo, Copainalá Zoque, Kosarek Yale, Miskito, Manange, Embera, Cayapa, Rama, Bislama, Mbum, Hupda,	"Isolating" languages (§ 4.6.1) Compositional type of polysynthesis (§ 4.6.2) North American languages with verb-centered nominals (§ 4.6.4.) Mixed languages of Western Eurasia (§ 4.6.5)
Quartile IV:	Lexical type dominant, derivational type subsidiary or unattested	Abzakh Adyghe, Cheyenne, Ket, Yir Yoront, Katcha, Ngambay, Baruya, Berik, Kaluli, Kwoma, Meyah, Toaripi, Badaga, Nivkh, Kildin Saami, Highland Chontal, Wappo, Bwe Karen, White Hmong, Yay, Buli, Dongolese Nubian, Rendille, Kyaka, Mandarin, Vietnamese	"Isolating" Languages of SE Asia, the Americas, Africa and New Guinea (§ 4.5.1) Compositional Type of polysynthesis (§ 4.5.2)

table 1: relative occurrence of the derived and lexical type of complex lexemes in the languages of the core sample, divided into quartiles

It is this embryonic typology that will be the starting point for the further discussion, and as the discussion proceeds step by step through the analysis of the data here and in chapter 5, further typological correlations and dimensions will be added to the classification.

A word of caution: since this rough distinction is based exclusively on the data for the 160-meaning wordlist, the characterization is obviously restricted to the occurrence of

either formal relation in this part of the vocabulary, and need not necessarily coincide with the outcome one would obtain when considering a more inclusive set of lexical items.

4.3. AREAL DISTRIBUTION OF THE TYPES

In addition to establishing typological groupings, the areal distribution of the types is highly uneven cross-linguistically, to the effect that such a classification also leads to the emergence of areal types. A map on the basis of the core sample showing the areal distribution of the relative degree of lexical items of the derived type is seen in figure 2. The size of the dots represents the frequency of occurrence of terms of the derived type: the larger the dots, the more prevalent derived terms are among the analyzable terms (similar coding by size will be used for the visualization of other quantitative variables in subsequent chapters).

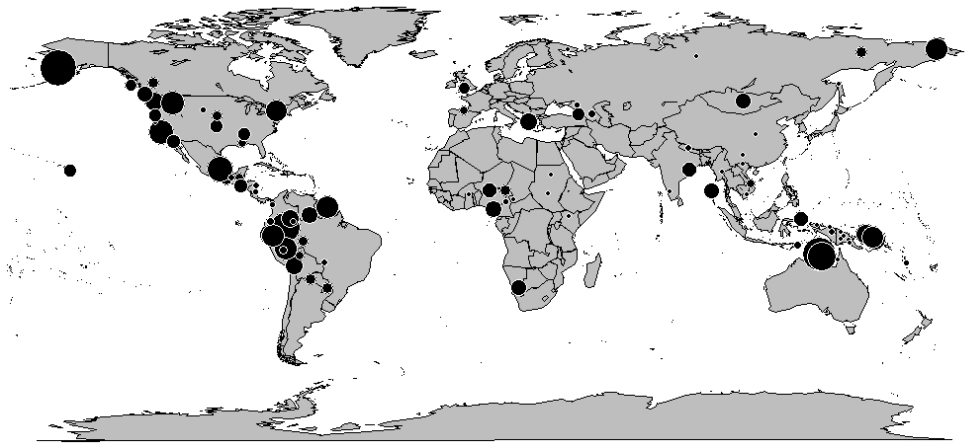

fig. 2: relative degree of lexical items of the derived type, core sample

However, it is a matter of debate and current discussion as to how areality in the distribution of linguistic features is best assessed and controlled for in typology, a problem which will be more extensively discussed when it comes to lexico-semantic patterns that appear to be areal phenomena in § 6.4.3. For the broad analysis in terms of the quantity of the main variables surveyed, as well as to avoid defining areas in a mere ad-hoc fashion that would potentially lead to the artificial emergence of areality, three different established breakdowns of the world into rather large linguistic areas are used throughout. These divisions have been defined independently of the present study to assess areality for its own sake or for quite different purposes. They are:

(i) DRYER'S (1992) SIX-AREA BREAKDOWN (Dryer-6 for short): a division of the world into six linguistic macro-areas: Africa, Eurasia, Australia-New Guinea, North America, South America, Southeast Asia and Oceania.

(ii) NICHOLS'S (1992: 25-26) ELEVEN-AREA BREAKDOWN, i.e. with Europe separated from Eurasia and Western from Eastern North America as in some of her analyses (Nichols-11 for short): Africa, New Guinea, Oceania, South and Southeast Asia, Australia, Europe, Eurasia, Western North America, Eastern North America, Mesoamerica and South America[2]

(iii) NICHOLS'S (1992: 27) THREE-WAY BREAKDOWN (Nichols-3 for short) into Old World, New World and the Pacific. The Old World comprises Africa and Eurasia (including Southeast Asia), the Pacific region is made up of New Guinea, Oceania and Australia, while the New World is constituted by the Americas

Under all three breakdowns, the areal distribution of the differential degree to which languages rely more on the derived and the lexical type is highly uneven and statistically significant (Dryer-6: $\chi^2 = 11.5158$, df = 5, $p = .04206$, Nichols-11: $\chi^2 = 23.9079$, df = 10, $p = .007849$, Nichols-3: $\chi^2 = 10.793$, df = 2, $p = .004532$, Kruskal Wallis rank sum tests). This and all further statistical analyses were carried out using the R statistical computing environment (R development core team 2009). Figures 3 to 5 plot the results for the different areal breakdowns.[3]

[2] Nichols's (1992) Ancient Near East is omitted for the obvious reason that no relevant language is present in the current sample.

[3] The plots are so-called boxplots, which are a useful way to visualize differences with respect to a variable (here: the percentage of terms of the derived type) between certain groups (here: languages of different areas of the world). The two most important types of information that can be extracted from a boxplot are: (i) the mean of the values within one group, which is represented by the thick black band in the middle of the boxes, and (ii) the variance around that mean. Specifically, the height of the boxes and the length of the whiskers (the dashed lines at both ends of the boxes) provide information as to this variance: large boxes (the size of the boxes includes 50% of the observed values) and long whiskers (for the plots shown in this thesis, whiskers extend to up to those datapoints no more than 1 time the interquartile range removed from the box) indicate high variance around the mean, while small boxes and short whiskers (or no whiskers) indicate that the observed values are grouped closely to the mean. The width of the boxes corresponds to the number of observations within each group: wide boxes show that there are many observations for that group, while relatively narrow boxes indicate that there are correspondingly fewer observations within the group. Thus, in the plots in figure 3 to 5, the fact that the boxes for the Americas are wider than those for Africa and Southeast Asia corresponds to the higher genetic diversity and hence more sampled languages for this area.

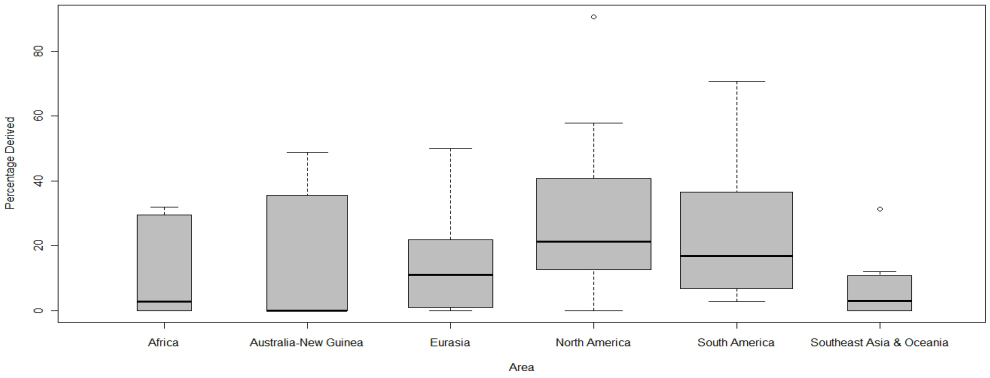

fig. 3: areal breakdown of the occurrence of terms of the derived type, using Dryer's (1992) breakdown

Here there is a lot of variation, but on average elevated degrees of derived terms in the Americas and rather low values in Africa, Australia-New Guinea, and Southeast Asia and Oceania. The Eurasian area falls somewhere in between (compare § 4.6.6.), but tends more towards low ratios of the derived type as found in Africa and Southeast Asia and Oceania. The more fine-grained Nichols-11 breakdown reveals that the distribution within North America is quite uneven, with higher degrees found in the west than in the east. Further, it reveals a difference between New Guinea and Oceania, with the former area having very few and the latter quite a lot of derived terms.

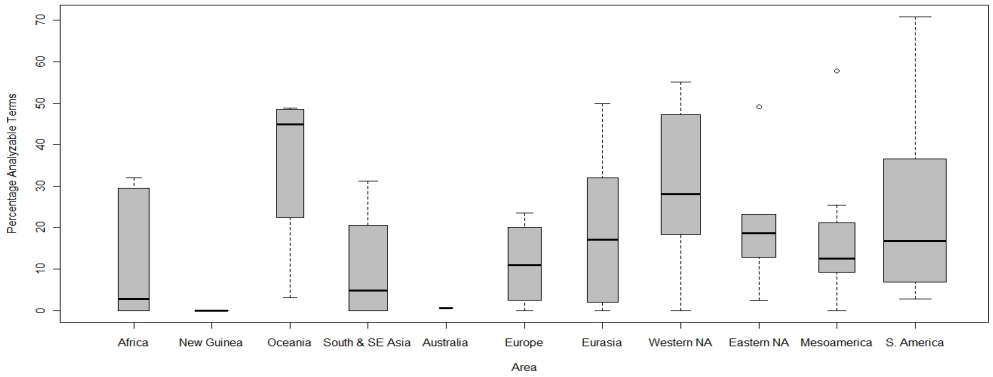

fig. 4: areal breakdown of the occurrence of terms of the derived type, using Nichols's (1992: 25-26) breakdown

Conflating these areas into the three very large areas of the Nichols-3 breakdown, one can observe a general upward trend as one moves to the New World, where the derived type is most frequent.

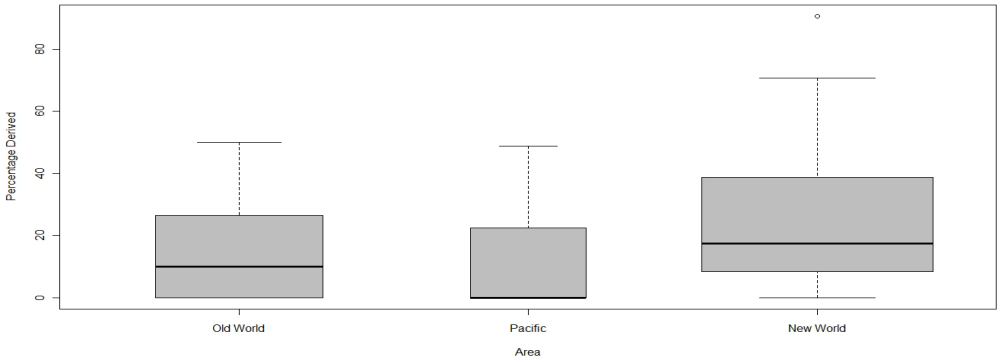

fig. 5: areal breakdown of the occurrence of terms of the derived type, using Nichols's (1992: 26-27) breakdown

4.4. DERIVED TYPE DOMINANT

4.4.1. DERIVATIONAL USE OF NOUN CLASSIFICATION SYSTEMS

4.4.1.1. Northwest Amazonia

This section is concerned with languages of South America, more exactly the Northwest Amazon area, with a profile rich in derived terms which are due to the extensive utilization of their systems of nominal classification for word-formation purposes.

In Bora and Cubeo (as well as other Tucanoan languages), there are affixal noun classifiers which occur both in the context of numerals and demonstratives, and which cannot be incorporated into the verbal complex. There is a boundary between lexical nouns and classifiers which is at times somewhat fluid; the latter may correspond partially or completely to noun roots in their segmental shape (Payne 1987: 26). This is likely due to the recent grammaticalization of classifiers from free-standing nouns (Aikhenvald 2000: ch13); the correspondence between classifier and lexical noun root is also apparent in Chayahuita, compare *-i'*, the Chayahuita classifier for liquids and the identical *i'* 'river, water.'

There is some variation in the properties of classifier systems in the Amazon discussed in Payne (1987) and Derbyshire and Payne (1990). For instance, the difference between Chayahuita and languages like Bora and Tucanoan languages is that the former allows for incorporation of classifiers into the verb complex, as seen in (1.), while the latter do not:

(1.) *a'pë-raya-t-ër-in*
burn-CL:disc-TRANS-INDICATIVE-3:SUBJ
'(he) burned (his) face'[4] (Payne 1987: 34)

[4] Glosses: TRANS 'transitive,' SUBJ 'subject.'

Furthermore, classifiers in Arabela differ from those of Huitotoan and Tucanoan languages in that they do not occur in the context of numerals (of which there are only three native ones) and accompanying demonstratives (Payne 1987: 29). Classifier systems are absent in Carib, Gê, and Tupi-Guaraní languages (although some Cariban languages may feature incipient noun classification systems, the same is true of the Tupian language Munduruku, according to Gonçalves 1987 as cited in Aikhenvald 2000). In Arabela, Chayahuita, as well as Bora and Cubeo (and other languages of their respective families), classifiers may function as nominalizers (Derbyshire and Payne 1990: 253, 266).

In fact, the first point to make is that noun classification devices provide a powerful means of word-formation that is used to enrich the lexicon. This is an aspect of nominal classification that is apparent in the set of vocabulary items under investigation here, and pertains noticeably the linguistic integration of items of acculturation, but not exclusively so.[5] However, classifiers functioning as nominalizers do not seem to be the whole story; they apparently also serve as derivational devices in general, and are not necessarily restricted to the application to verbal roots (as noted by Payne 1987: 33 for Chayahuita specifically). Noun classifiers are also present to some extent in Yanomámi. There is a class of "characteristic classifiers," which have the typical semantics expected in classifier systems: *koko* 'round/cylindrical,' *ma* 'hard,' *si* 'thin and flat,' and *mo* 'small and round' (Goodwin Gómez 2000: 15, describing a slightly divergent dialect than that of the consulted source). The origin in lexical nouns is often recognizable (such as *mo* < *mamo* 'eye, eyeball,' *si* < *si kɨ* 'skin'). Not all nouns are associated with characteristic classifiers, but most may be accompanied by so-called "quanta-specifying" classifiers. In particular, there are two such classifiers associated with plurality, *kɨ* and *pə* in Goodwin Gómez (2000) and *kɨ* and *pë* in the consulted lexical source. Semantic differences between them are hard to pin down (Goodwin Gómez 2000: 17). There is also the classifier *kɨkɨ*, apparently identical with *këkɨ* in the consulted source, glossed as 'collective' by Goodwin Gómez (2000: 16). Thus, for instance, *shi* is 'faeces' and *shi-kɨ* ~ *shi-pë* is 'guts,' *here* is 'soft,' and *here-kɨ* is 'lungs' (examples from database).

Examples from Arabela, Chayahuita, Bora and Cubeo for a selection of meanings amply illustrate a derivational use of classifiers.[6] In the slice of the lexicon investigated in the present study, derivational use of classifiers is more frequent in Bora and Cubeo, while it is found in a smaller number of examples in Arabela and Chayahuita. Perhaps this is related in some way to the typological differences in the classifier system between these languages and Bora and Cubeo as representatives of Huitotoan and Tucanoan respectively, as summarized above from Payne (1987) and Derbyshire and Payne (1990). The examples

[5] Compare also Seifart's (2005: 209fn25) statement about the Miraña dialect of Bora: "Where at all possible, Miraña favors the creation of neologisms over borrowing of morphological material. In the case of concrete objects, in particular artifacts, class markers play a major role in these creations" as well as Morse and Maxwell (1999: 73), who state that the Cubeo classifier system "enables a Cubeo who sees an unfamiliar object for the first time to give it a Cubeo name."

[6] Glosses for classifiers have been created in an ad-hoc fashion for the present study specifically, with the exception of Bora, for which Seifart's (2005) conventions are followed and for examples incorporated from sources other than the primary ones consulted.

also show that the morpho-semantic structure of the relevant lexical items is often very similar.

(2.) 'milk'
a. Arabela *quitia-aca* 'breast/teat-CLASS.LIQUID'
b. Chayahuita *sho'sho-i'* 'breast-CLASS.LIQUID'
c. Bora *mújpañejpácyo* /mújpañe-hpacyo/ 'breast-SCM.LIQUID'

(3.) 'table'
a. Bora *méétsa-wa* 'table-CL.plank'[7]
b. Cubeo *tʉoiva* /tʉoyʉ-va/
'serve.food-CLASS.BROAD.AND.FLAT.OBJECT'

(4.) 'ladder'
a. Arabela *taka-tu* 'go.up-CL'[8]
b. Bora *neríívye-wááhyo* 'ascend-CL.layered.things'
c. Cubeo *mʉi-caðava* 'ladder-CLASS.FRAMELIKE.GRIDLIKE.OBJECT'[9]

(5.) 'tear'
a. Arabela *namiji-aca* 'eye-CLASS.LIQUID'
b. Chayahuita *na'nëi'* /na'nërin-i'/ 'cry-CLASS.LIQUID'
c. Bora *máátyo-u* 'crying-CL.ROUND'
d. Cubeo *ori-yaco-rʉ* 'tear-eye-CLASS.ROUNDISH.OBJECT'

(6.) 'honey'
a. Chayahuita *nino-i'* 'beehive-CLASS.LIQUID'
b. Bora *íímúhojpácyo* /íímúho-hpácyo/ 'beehive-CL.LIQUID'
c. Cubeo *mumicoro* /mumi-córo/ 'bee/honey-CLASS.LIQUID.STATE'

(7.) 'cave'
a. Chayahuita *na'pi-të-ana* 'rock-??-CLASS.AROUND'
b. Bora *ííñuhéju* /íñu-héju/ 'earth-CL.HOLE'
c. Cubeo *cʉ̃ra-cobe* 'stone-CLASS.HOLE-LIKE.OBJECT'

[7] Note that in Bora, the word for 'table' is apparently a loan from Spanish (< *mesa*) integrated into the Bora lexicon with a suffixed classifier.
[8] From Payne (1987: 30), citing Edgar Pastor (p.c.).
[9] The Cubeo example for 'ladder' illustrates that the presence of the classifier morpheme is at times redundant or seemingly redundant on the level of the lexical item, since in this case *mʉi* alone already denotes 'ladder.'

(8.) 'nostril'
a. Arabela *naju-hua* '??-CLASS.HOLE'[10]
b. Chayahuita *nitëana* /nitë'-ana/ 'nose-CLASS.AROUND'
c. Bora *túúheju* /túú-ʔehɯ/ 'nose-SCM.HOLE'[11]
d. Cubeo *ũe-cobe* 'nose/nostril-CLASS.HOLE-LIKE.OBJECT'

While the emergence of noun classification in unrelated languages of the Amazon itself may be due to language contact (Payne 1987: 37-38, Seifart and Payne 2007: 384-385), this need not necessarily be so in the case of the morpho-semantic similarities notable in some of the above examples, although at least in some cases areal spread through language contact appears at least possible. One case is to have terms for 'tear' based on a verb root for 'cry,' which is not very common cross-linguistically (compare Appendix E, 140). Also, the fact that Bora and Chayahuita derive their terms for 'honey' from 'beehive' rather than 'bee' is not attested in other regions of the world, judging from the evidence of the sample (see Apendix E, 20).

Cubeo has more than 150 classifiers (Morse and Maxwell 1999: 75), which is one of the largest systems known, and there is evidence that they, at least in this language, also serve as singularizers, with (some of the) roots they attach to denoting masses rather than individualized entities themselves (see Lucy 1992 on this aspect of nominal classification). Thus, for instance, Cubeo *nʉo* means 'roots' generically, and *nʉo-mu* 'roots-CLASS.LIANA-LIKE' is used for a single root, *ori* denotes 'tears' collectively, and *ori-yaco-rʉ* 'tears-eye-CLASS.ROUNDISH.OBJECT' (compare 5d.) is used for a single tear. Another language spoken in the relative vicinity that has been claimed to feature a system of noun classification is Rama, of the Chibchan family. Craig (1990), however, prefers to speak of class terms, which have semantically common with classifiers that they categorize nouns depending on their shape and consistency. There are four such class terms that she accepts, three of which have a free noun counterpart: *kat* 'tree/long rigid object,' *up* 'eye, seed/round, solid object,' *kaas* 'flesh, meat/fleshy object,' and *li ~ ri*, which has no lexical counterpart and is used in association with liquids. However, Craig notes explicitly that, given that there are lexical counterparts for three items on the above list, it is difficult to distinguish between ordinary noun compounding (which is also widely used in the language involving other elements than those above) and nominal classification, and in fact does not mention any structural criteria that would allow one to do so, concluding that the phenomenon is relatively marginal in the grammar of the language.

4.4.1.2. Australia and Oceania

There are also languages in the sample outside the Northwest Amazon that employ a system of nominal classification for derivational purposes, clustering in Australia and Oceania. One of them is Burarra, spoken in Australia. In fact, the use of classifiers, which in this case come in the form of prefixes, is the prime mechanism by which the relatively few morphologically complex terms for the investigated meanings are formed in Burarra.

[10] Compare *najacua* 'nose.'
[11] Meaning of *túú* not entirely sure.

There are four noun classes, 'domestic,' marked by *mun-*, 'general,' marked by *gun-*, 'masculine,' marked by *an-*, and 'feminine,' marked by *jin-* (class labels from Glasgow 1994). Green (1987: 15) notes difficulties in figuring out a semantic basis for noun class assignment. According to Green (1987: 15), noun class is obligatorily marked on adjectives and verbs, depending on the grammatical context. In contrast, noun class is not overtly marked for most nouns (there are some nouns with a noun class prefix forming part of the stem). However, there are nouns which have variable noun classes, and these are the ones of interest here, since the semantics are altered depending on which class prefix is chosen. Examples include:

(9.) a. *mun-ngarnama* 'CLASS.DOMESTIC-inner.thigh' = 'bark'
b. *gun-ngarnama* 'CLASS.GENERAL-inner.thigh' = 'cave'
c. *gun-gapulawa* 'CLASS.GENERAL-mound.up' = 'nest'
d. *an-giya* 'CLASS.MASCULINE-egg' = 'calf'

Examples (9a.) and (9b.) not only illustrate the pervasiveness of transferring body-part terms to topological features in Australian languages (noted for instance by Schebeck 1978 and Dixon 2002: 99), but also that the classifier system provides means to derive several different meanings from the same lexical root by assigning it to different noun classes. (9c.), in contrast, illustrates an apparently purely nominalizing usage of the 'general' classifier. As for other languages of Australia in the sample, derivational use is also attested in Nunggubuyu, but appears to be much less frequent. Gurindji does not feature noun classes, and Yir Yoront is part of an area on the western side of the Cape York Peninsula where languages instead feature a productive system of compounding, involving body-part terms in particular (Dixon 1980: 111).[12]

Noun classes are also a feature of Rotokas, spoken on the island of Bougainville (see Terrill 2002 for an overview of noun classification systems in Oceania more generally). Rotokas nominal inflectional morphology is not very elaborate. There are no core cases; semantic relations otherwise susceptible to being expressed by case (e.g. locative, allative, benefactive, ablative) are coded by enclitics. Nouns are optionally inflected for number and gender, and verbs agree with the A of transitive clauses, but not with the P (see Robinson 2011: 79-92 for full discussion of Rotokas nominal morphology). Rotokas has an apparently closed set of four shape-based classifiers: *isi* 'round object,' *kuio* 'round object (edible),' *ua* 'narrow object,' and *kae* 'long object' (Robinson 2011: 50), as well as a number of classifiers which are not shape-based but instead are used for collections of certain semantic types, such as *kou* 'heap' (Robinson 2011: 51).[13] For derivational purposes, the shape-

[12] In fact, Dixon (1980: 111) notes that "many compounds would be translated by simple roots in languages from other regions" so that Yir Yoront is probably a fairly atypical language in the broader areal context of Australia, and this may be the reason for the low occurrence of the derived type in the areal evaluation above for Australia as represented by Yir Yoront. Indeed, the language has the highest percentage of morphologically complex terms of all sampled languages in Australia, although it is closely followed by Burarra. In contrast, the ratio of morphologically complex terms in Nunggubuyu, Ngaanyatjarra and Gurindji is drastically lower.

[13] Robinson (2011: 51) states that classifiers are a separate word class in Rotokas rather than being bound as suggested in earlier analyses, which is why they are written here as separate words.

based classifiers are relevant, as seen in the examples in (10.). They occur in all investigated semantic domains expect for the fourth (phases of the day and miscellanea), though they are most common in the domain of nature-related and topological terms:

(10.) a. *kaki ua* 'cracked[14] CLASS.NARROW.OBJECT' = 'cave'
b. *vuvui ua* 'transparent CLASS.NARROW.OBJECT' = 'sky'
c. *ripo kae* 'cocoa.husk CLASS.LONG.OBJECT' = 'eyelash'
d. *vovou isi* 'breathe/want CLASS.ROUND.OBJECT' = 'heart'

Some languages associated with the derived type feature other dedicated derivational morphemes in addition to noun classification devices. Here, the overall effect is thus due to a combination of the presence of noun classification employed in a derivational fashion on top of "ordinary" derivation. This is the case in Rotokas, which has the general derivational morpheme *-pa* to derive instrument and agent nouns, and exploits this resource along with its system of nominal classification in creating lexical items. It can be seen in action twice in (11.).

(11.) *ori-pa-to* *Raka* *eisi ruvaru-pa* *kepa*
cook-DERIV-SG.M Raka LOC heal-DERIV house
'Raka is the cook at the medical station' (Robinson 2011: 83)

The suffix *-pa* is particularly frequent in terms for items of acculturation, such as *ravireo vetaveta-pa* 'sun count-DERIV' = 'clock.'[15]

Mali, another language of Oceania, also has a system of nominal classification which is exploited for lexical expansion (Stebbins 2005: 102). Thus, the unmarked form of the root *isem* 'bird' occurs with the masculine noun class suffix *-ka* (*isem-ka* 'a bird') while, with the feminine noun class suffix *-ki*, it means 'airplane.' Similarly, compare *vilē-ka* 'crack of lightning' with *vilē-ki* 'gun,' and *slēp-ki* 'bone' with *slēp-ka* 'pen, pencil' (the feminine noun class is associated with big size in Mali according to Stebbins 2005; this is also true of many other languages, see § 6.2.3.4. for discussion). However, the classification system is also exploited to vary the semantics of native lexical morphemes. Among the examples provided are the following forms of the root *amēng* 'wood' (Stebbins 2005: 86, table 4): *amēng-ka* 'slender tree' (masculine class), *amēng-ki* 'large full grown tree' (feminine class), *amēng-ini* 'stick' (diminutive class), *amēng-ēm* 'tree stump' (reduced class), *amēng-igl* 'plank' (excised class), *amēng-vēt* 'pole' (long class), and *amēng-ia* 'large log' (extended class).

A system of nominal classification or noun classes might also be a feature of Great Andamanese, although the situation remains unclear on the basis of the available evidence (also note that Great Andamanese is in the second quartile of the division made above, but on the top edge). Manoharan (1989: 61) says that Great Andamanese "uses a stem or root

[14] Full gloss: 'opened, cracked, split open.'
[15] At times, gender/number suffixes appear to be employed in a derivational fashion as well, such as in *kaporo-to* 'space-M.SG' 'scissors,' but this type of formation remains unclear.

for one concept" and then may vary the semantics of that concept "differently using different affixes what [sic!] are called 'formative affixes'." Examples include (adapted from Manoharan's 1989: 61-63 more extensive list):

(12.) a. *-beːc:*
i-beːc 'honey comb'
araː-beːc 'tail (of birds)'
tɔt-beːc 'feather'
ʈɔt-beːc 'hair-my'
ʈɔw-ter-beːc 'cloud' (*ʈɔːw* 'sky')
ṅeruluto-beːc 'eyelash-your' (*ṅe-erulu* 'eye-your')
ṅertaːp-beːc 'beard-your' (*ṅe-ertaːp* 'your-chin')[16]
taːjiyo-tot beːc 'bird' (*ta jiyo* 'pool')

b. *ɸoṅ* 'hole, cave'
myoː-ɸoṅ 'cave' (*myoː* 'rock')
inɔ-tɛr-ɸoṅ 'well' (*inɔ-* 'water')
ṅerkɔːtho taraː-ɸoṅ 'nostril-your' (*ṅerkɔːtho* 'nose-your')
ṅa-ɸoṅ 'mouth-your' (*ṅa-* 'your')[17]
ɲɔm-ɸoṅ 'armpit'

Burenhult (1996: 10) calls these "formative prefixes," but from the examples in Manoharan (1989: 61-64) it seems that the recurrent elements in fact more often, but not exclusively, follow rather than precede other material. Given the variability in position extractable from the examples, it is unclear whether one is dealing here with something other than "ordinary" compounding. A system of body part classification that is extended to the domain of kinship is mentioned by Abbi (2006b). Since the majority of examples do not in fact involve body-parts as the putatively classified entity, it is unclear how the system of these affixes (if it is a system) is related to the system of body-part classification outlined in Abbi (2006b); Manoharan (1986: 28) speaks of two different "types," which suggests that they are independent of one another. There are also similar prefixes in semantically related terms: compare for instance *âkà-bang* 'mouth' with *âkà-pai* and *âkà-pē* 'lip' and *âkà-êkib* 'jaw,' *ông-pâg* 'foot' with *ông-tôgur* 'ankle' etc. To add to the confusion, these prefixes are described as indicating the possessor of body parts in Man (1923), although they clearly also occur on terms not denoting body-parts. Further, the shape of the prefixes does not match those mentioned by Abbi (2006), although again it may be relevant that present-day Great Andamanese is a conglomerate of several erstwhile distinct

[16] Compare Abbi (2006a: 31): *ʈhɛr-ʈap'-bec'* 'my beard'

[17] But compare Abbi (2006a: 31) *tʰɛr-ɸoŋ* 'my mouth.' *ʈʰa*, which should correspond orthographically to *ṅa-* in Manoharan is used for 'tongue' and kinship terminology according to her, though note that there is variation between speakers and that "[c]onsistency is lacking, as the language contains several varieties of past and present dialects. In general, a speaker has two or more forms in the verbal repertoire that s/he can vary freely in all contexts ..." (Abbi 2006b: 114).

varieties or languages (Manoharan 1986: 27-29 reports a wide range of allomorphy in present-day Andamanese resulting from this mixture, and provides the following list of attested prefixes: *it-*, *et-*, *ir-*, *er-*, *akka: ~ ekka:-*, *ara: ~ era:-*, *atta: ~ etta:-*, *ot-*, *tot-*, *ta:-*, *tara:-*, *i-*, *0-*). The discussion in Manoharan (1986: 28) suggests an extremely interesting system of classification based on conceptual dependency. As his example goes, 'pig' is independent, but 'head' is not in that it is relational and attached to a larger entity, and thus the latter receives marking with a "classifier." Compare *mōlo* 'egg' from the database with *âr-mōl* 'bud' (this association is also found elsewhere, see Appendix E, 9), which would fit nicely into this scheme: an 'egg' is conceptually an independent unit, since it can be picked up and carried away (it is detached, in terms of Gibson 1979), while a 'bud' is attached to a larger structure, namely the budding plant, and as such does not enjoy the same degree of conceptual independence and is marked by a formative prefix attached to the same lexical root. Manoharan (1986: 30) also emphasizes that dialect mixture may be responsible for blurring an erstwhile more straightforwardly semantically based system of formative prefixes assignment, with vestigial traces of the putative original system left. For instance, the marker *ara: ~ era:-* is found in many names of objects at the back of the larger structure (e.g. *ara:-be:c* 'tail of bird,' *ara:-ce* 'sting,' *era:-bu:-co* 'roots'); note also the transition in Bantu from a putatively initially semantically based to a formally based system of assignment (Corbett 1991: 48-49). Manoharan (1986: 30-31) also attaches importance to the appraisal that the Andamanese system is different from the classifiers found in languages of mainland India: "they are not here called classifiers since the primary function of these formative affixes is to change the meaning of the concept attached to the root and not to classify the nouns or that of the kind" and summarizes: "in the case of formative affixes, the language does not possess different words for different concepts; rather, it changes the meaning of the primary concept into the specific meaning." Regardless of the precise properties of the Andamanese system which is not entirely clear on the basis of the available evidence, this is precisely the aspect which is relevant for the present discussion.[18]

4.4.1.3. Classifiers in Southeast Asia

Classifiers are also present in Southeast Asian languages.[19] In line with the isolating typological profiles of these languages, they are typically independent words in these languages rather than affixal, as is predominantly the case in the Amazon region (see Aikhenvald 2007: 10). Burmese data from Becker (1975) presented in Aikhenvald (2007: 11) indicate that the numeral classifiers in Southeast Asian languages may be used to variegate the lexical semantics of nouns in discourse and to highlight aspects of their meaning (see also Bisang 1999: 129-130). However, classifiers in Southeast Asian languages appear to play a much smaller role as derivational devices and thus as a means to enrich the lexicon when compared with the situation in the Amazon; Bisang (1999) surveys functions of classifiers in a number of Southeast Asian languages, but does not mention derivational

[18] Compare Burenhult (1996: 10): "[i]t is extremely difficult to get a clear idea of this system of formative prefixes, and the lack of raw data prevents us from taking the analysis any further..."

[19] There are also entities called "verbal classifiers" in languages of North America, for instance Haida, which serve to specify the semantics of the verb they attach to; these elements fall outside the scope of the present discussion because of being associated with verbs without changing word-class.

usage. However, Bisang (1999: 114) mentions that the "high degree of indeterminateness of nouns is a very important characteristic of East and Southeast Asian languages ... which is crucial to the existence of classifiers." Still, this indeterminateness does not appear to be resolved at the level of the lexical item by using classifiers in a derivational fashion, but rather to modulate the semantics of the classified noun at the discourse-level (though see also § 4.5.1.3.2. for some evidence from Yay). Consequently, most Southeast Asian languages in the present sample are found in the fourth quartile, with terms of the derived type exceedingly rare (compare discussion in § 4.5.1.3.)

Moving back to the general discussion, it has to be pointed out that the sheer presence of a system of nominal classification does not necessarily entail its employment for derivational purposes. Nominal classification is also found in Meyah (Gravelle 2004) and yet, judging from the database, it does not appear to make use of its noun classification system on a large scale for word-formation. Here, differences in the individual system, such as whether classification occurs predominantly in the context of numerals, obviously may play a role.

4.4.1.4 Excursus: Noun Classes and Gender

Noun classes are also found in languages of Africa, most notably in languages of the Bantu subgroup of the Niger-Congo family, and here, the relevant markers are also employed in a derivational fashion. In terms of Corbett (1991, 2007), genders are identified on the basis of syntactic evidence, or, more precisely, by the ability of the gender of a head noun to trigger agreement on other elements of the sentence and thus to constitute agreement classes. In Bantu languages such as Swahili, gender is indicated by a prefix on noun roots. A complicating aspect is that prefixes on nouns do not always match the agreement forms (Corbett 1991: 44-45), a fact that makes it necessary to distinguish between agreement classes, which are relevant evidence for identifying gender, and inflectional classes of noun roots. It is the latter class which is relevant here, since indeed these inflectional classes may be used in a derivational fashion.

In Swahili, for instance, markers for inflectional class may be prefixed to a nominal or verbal root to form the names for artifacts (13a.), and in the domain of natural phenomena and body parts, the same root may be used with varying inflectional prefixes to alter and specify the semantics encoded in the root (see also Corbett 1991: 44), as in (13b.)

(13.) a. *u-funguo* 'key' (*fungua* 'to open')
b. *u-tumbo* 'guts' (*tumbo* 'belly, stomach')

Mufwene (1980) emphasizes that class prefixes of the Bantu type are not just inflectional, but also derivational, arguing on the basis of similar evidence to that presented here. Swahili, however, does not exclusively rely on noun class alternation as a motivating device, but frequently employs syntactic devices as well to create new lexemes in phrasal form, for instance *shavu la mguu* 'cheek of leg' = 'calf.'

Having moved the discussion to Bantu-type noun class systems, a review of gender is not far away. In fact, noun class systems are frequently subsumed under the notion of gender, for instance by Corbett (1991), who distinguishes between semantic and formal

systems of gender assignment and discusses noun classes in Bantu languages as being of the formal kind (although they have a semantic component and may have started out diachronically as being assigned on a semantic basis alone, Corbett 1991: 48-49). Furthermore, gender systems are sometimes employed to alter the semantics of lexical items by assigning the same lexical item to different genders, for instance, German *die See* (feminine) 'sea,' *der See* (masculine) 'lake.' Scattered examples of this phenomenon are found in a number of languages in the sample. However, one language that frequently utilizes variation in the assignment to genders of the same lexical items as a motivational device is the Choco language Embera, as seen in the examples in (14.) (glossing is somewhat simplified here for ease of presentation, and some of the stated meanings in the source are omitted).

(14.) a. *hú* 'breast, chest' (masc), 'teat, nipple' (fem.)
b. *kẽŋgú* 'nerve' (masc.), 'vein, lode' (fem.)
c. *kathárro* 'catarrh' (masc.), 'phlegm' (neut.), 'snot' (fem.)
d. *káarta* 'paper' (masc.), 'letter' (fem.)

As examples (14c.) and (14d.) show, variability in gender assignment is also present in loanwords from Spanish.

Moving to a general discussion of noun classification from the perspective of lexical typology specifically, an important aspect of derivation by noun classification is noted by Aikhenvald (2000: 267): "Noun classifiers are often used to highlight different aspects of the meaning of a polysemous noun." This can be nicely illustrated with examples from Cubeo. For instance, the root *cu̱ra* in (6c.) only occurs suffixed with classifiers to the effect that its semantics can in fact only be extracted by comparing that of the different terms with classifier and "subtracting" their respective meaning. As another example from the consulted source, the root *ũe* in fact only occurs suffixed with classifiers. With *-ca* 'tuber-shaped,' the derived form means 'nose, snout,' with *-bo* 'round or hard,' it means 'nose of an airplane,' and with *-cobe*, as in the above example, 'nostril.' Thus, "[t]he ways the presence of classifiers in a language can reduce the actual lexicon by expressing the same concepts through grammar relates to a well-known facet of the interface between grammar and lexicon" (Aikhenvald 2000: 268). [20] From the point of view of lexical typology, another property of the languages with noun classification devices in the sample, in particular those of the Northwest Amazon, is also noteworthy, namely that they lack patterns of colexification that are frequent elsewhere. This is particularly noticeable for the meaning pairs 'milk' and 'breast' (in all the languages surveyed above 'milk' is derived from 'breast' by means of a classifier) as well as the lack of colexification of 'bee' and 'honey,' which is a less frequent, but still common enough phenomenon cross-linguistically (see Appendix E, 30 and 123). It seems that the presence of a classifier system leads these languages to favor expressing these meaning pairs by derivation rather than colexification. This would need to be assessed against a larger number of classifier languages of the

[20] As Frank Seifart (p.c.) points out, this can only be said to the extent that classifiers actually are grammar. As noted in the above discussion, many classifiers can be traced back diachronically to free-standing nominal roots, and in some cases, they are in fact identical in shape to nouns.

Northwest Amazon, but there does appear to be a (weak) correlation, although probably no strict law, to this effect.

4.4.2. AFFIXAL TYPE OF POLYSYNTHESIS

This type is mostly found in the American Northwest, and is chiefly responsible for the high frequency in derived terms there. It is most clearly represented by Central Yup'ik, an Eskimo-Aleut language spoken in Alaska. Like other Eskimo-Aleut languages, it is notable for the absence of compounding, employing only derivational mechanisms for word-formation purposes (Mather et al. 2002: 16, although Mithun 2009: 13 argues that noun-suffix structures in Eskimo-Aleut languages have recently originated from incorporation-like structures). Derivation in Central Yup'ik operates by a class of morphemes called postbases in Eskimo-Aleut studies. These are exclusively suffixal, and may be applied iteratively, that is, a lexical root (and note that the defining feature of this type in the present study is that only one lexical root be involved!) may be suffixed with several postbases that add to or modify their meaning. According to Mather et al. (2002: 33), postbases in Central Yup'ik can be classified according to whether they are trans-categorical or not, that is, whether they change the lexical class of the root (or of the root already modified by postbases, which is called an expanded base) they attach to (verbalizing/nominalizing postbases) or whether they preserve the lexical class of the root (or expanded base) they attach to (nominal-elaborating postbases/verb-elaborating postbases).

The following construction (from Mather et al. 2002: 35, glosses inferred) illustrates both types:

(15.) *teng-ssuu-ksuar-Ø*
fly-instrument-small-ABS
'small plane (flying instrument)'

Teng is a verbal root meaning 'to fly,' *-ssuu* is a nominalizing postbase that creates instrument nouns, and the following *-ksuar* 'small' is a nominal-elaborating postbase adding the meaning 'small' to the preceding material. Finally, there is a zero-marked absolutive case marker (nominals are obligatorily inflected for case). As the example also shows, postbases may be used in an ad-hoc fashion to modify the semantics of a base or expanded base, but they are also the prime mechanism responsible for the formation of institutionalized morphologically complex expressions in all semantic domains under investigation here (Fortescue 1980: 259, in discussing the related language West Greenlandic, reports frequent conventionalization of combinations of root and postbase and notes in general that in polysynthetic languages typically "the area of overlap between clearly lexicalized items and productive rules which can generate them from more basic morphemes results in an extensive descriptive 'penumbra.'"). Examples from the database (some of the terms only occur in some dialects of Central Yup'ik) include:

(16.) a. *yaqulek* /yaquk-lek/
'wing-one.having' = 'bird, duck, fowl, angel'
b. *merr'aq* /meq-rraq/ 'fresh.water-a.little.bit.of' = 'dew'
c. *agluryak* /agluq-yak/
'center.beam.of.a.structure-thing.similar.to' = 'rainbow'
d. *mingqun* /mingqe-(u)n/ 'sew-device.for' = 'needle'
e. *uquryak* /uquq-yak/ 'oil-thing.similar.to' = 'heavy sweat'

Note that by means of the postbase *-yak* 'thing similar to' the grammar of Central Yup'ik has a "built-in mechanism" to form metaphor-driven denominations. Still, contiguity dominates as the device underlying the motivated lexical items of Central Yup'ik. Note also that, in spite of the lack of means to combine lexical roots in the literature, there are some examples in the database that involve the combination or at least loose juxtaposition of two lexical roots such as *agyam anaa* 'star faeces' = 'meteor;' however, these are clearly marginal.

Languages which have a relatively high degree of analyzable lexical items of the derived type are common in the Americas, among them Nuuchahnulth (lacking both compounding and incorporation according to Davidson 2002: 92, although there are a number of examples of the database in which *mis* 'thing,' which has the potential to occur as a free-standing form, acts as a nominalizer), Nez Perce, Upper Chehalis, Ineseño Chumash in the Northwest and Xicotepec de Juárez Totonac in Mesoamerica. Some examples from Ineseño Chumash are in (17.).

(17.) a. *š-oqyokonič*
'3SG/3SG.POSS-be.uneven/protruding' = 'Adam's Apple'
b. *is-xɨp* 'one's.own-stone' = 'testicle'
c. *'aqmilimu'* /'aqmil-mu'/ 'drink-DERIV' = 'spring, drinking place'
d. *s-axiyi'* '3SG/3SG.POSS-be.dark' = 'night'

Derivation also plays a major role in word-formation in Chukchi, a polysynthetic language of Eurasia (see Dunn 1999: 143-148 for an overview of derivational devices). Complex lexemes of the derived type include:

(18.) a. *təlan* /təle-n/ 'go-LOC' = 'path' (see also Dunn 1999: 146)
b. *ajmə-n* 'fetch.water-LOC' = 'hole in ice, well'
c. *ine-newntet-icɣən* 'ANTIPASS-open-INSTR' = 'key, opener'[21]

However, Chukchi also allows for the concatenation of a variety of form classes into a nominal word form (see Mattissen 2003: 255-256 and Spencer 1995) and is consequently

[21] Verb stems must be intransitive for the instrumental suffix to be applicable; transitive verbs therefore must undergo antipassivization (or incorporation) first (Dunn 1999: 147); also note that the shape of the suffix is given as *-ineŋ(e-)* by Dunn (1999:147).

typologized as belonging to the compositional type of polysynthetic languages by Mattissen. Indeed, formations of the derived type are relatively rare in Chukchi when compared with Central Yup'ik or Nuuchahnulth (see brief discussion in 4.6.4.3.4.), although still more frequent than in most languages in the sample.

4.5. LEXICAL TYPE DOMINANT

4.5.1. ISOLATING TYPE

There is a discernible lexical profile of languages with predominantly isolating structure which involves analyzable terms of the lexical type as defined in § 3.6.1. (compare Aikhenvald 2007: 24, Vajda 2004: 400 and more tentatively Bauer 2009: 355 on the affinity of isolating languages with compounding). What it means to be an "isolating" language is, however, somewhat ill-defined. Following Greenberg (1960: 186), a relatively low ratio of inflectional morphemes per word may be taken as a hallmark of a prototypically isolating language; this is assessed here impressionistically for the time being, with reference to bound morphology in general, that is, inflectional morphology is frequently included in the discussions. Although inflectional morphology is not directly relevant for lexicalization processes and word-formation per se, the goal here is to show that the profile pertains to the morphological typology of the languages at large. This impressionistic assessment is to be systematized and operationalized in a bit more detail later. The isolating profile is illustrated with examples from languages from four different regions of the world: Africa, New Guinea, Southeast Asia, and the Americas.

4.5.1.1. Africa

4.5.1.1.1. *Ngambay.* Ngambay, a Nilo-Saharan language spoken in the Central African Republic, is a language in which all analyzable lexemes in the database are of the lexical type (see Moser 2004: 130 for the importance of compounding specifically). While Ngambay clearly features bound morphology, it is relatively sparse. For instance, there is no case marking for core grammatical relations, although there are inflectional categories such as inalienable possession as well as other non-bound case-like forms expressing relations such as dative and associative. Plurality is expressed by an independent word. Ngambay verbs may show agreement with pronominal subjects under certain conditions (Moser 2004: 290), and there is also a set of object suffixes (Moser 2004: 94). Analyzable terms in Ngambay range from the most common combination of two elements to phrasal idioms (still called compounds by Moser 2004: 132).

(19.) a. *ddéw-màann* 'way-water' = 'river'
b. *nɔn-kake* 'tear-tree' = 'resin'
c. *kàrè wùr énje* 'sun heart mother' = 'noon'
d. *bɔ̀-dìngàw wòy dɔ̀ ji-á té* 'big-man dies head arm-3S LOC' = 'sunset'
(Moser 2004: 132, small caps added)

Ngambay also features complex lexical items with *né* 'thing' such as *né kìsì* 'thing sit.down' = 'chair.' Moser (2004: 133) regards these as deverbal nominalization, and indeed, semantically they are equivalent to plain nominalization in languages with dedicated nominalizers. Since, however, *né* can also occur as a noun on its own, as shown by the example in (20.), such Ngambay nouns are considered to be of the lexical type here.

(20.) *gate né nèénn bàánn wà?*
price thing DEM how much QUE
'The price of this thing (is) how much?'[22] (Moser 2004: 300)

4.5.1.1.2. *Mbum.* Mbum, of the Adamawa branch of Niger-Congo, is a fairly isolating language, that is, the ratio of bound morphemes to free forms is low. The example in (21.), inferred from Hagege (1970: 209), illustrates some typical properties of this language type, such as lack of agreement on the verb and expression of plurality by an independent word rather than an affix.

(21.) *nzùk kà híŋà pínà Bí pàì rí*
man/human ASPECT do work in field PL
'the man/person is working in the fields'

Mbum has word-class changing derivation (Hagege 1970: 128-155), including processes to derive nominals. However, judging from Hagege's examples, the output of these processes are mostly abstract nouns, while the creation of terms for objects in the extra-linguistic world is largely taken over by complex lexemes of the lexical type (there is also one term in the data for the present study tentatively classified as being of the derived type, namely *ɓì-mbàm* 'at-rain' = 'puddle').

In the Mbum lexicon, there are basically two types of morphologically complex lexical items, both of the lexical type. The first involves the direct juxtaposition of two lexical roots which may be of different parts of speech (see Hagege 1970: 159-188 for more detailed discussion). Hagege calls the resulting structures compounds ("composes"), but gives no criteria concerning how to distinguish them from syntactic phrases. Some examples from the database are in (22.).

(22.) a. *ŋgàŋ-kpù* 'skin-trunk/tree' = 'bark'
b. *nûm-jóró* 'fat-bee' = 'honey'
c. *fè-nâm* 'thing-sleep' = 'bed, lit'

The second structural type, which is somewhat rarer in the database, consists of complex lexical items constructed using the connective morpheme *à* which is also found in Mbum possessive constructions (this marker is glossed as "juncteur" in Hagege 1970 and "connective" in Hagège 1993; the latter convention is used here). Examples are in (23.).

[22] Additional gloss: QUE 'Question.'

(23.) a. *pàk à jáù* 'house CONN urine' = 'bladder'
b. *pàk à ndòì* 'house CONN bird' = 'nest'
c. *sàk à mbàm* 'tear CONN rain' = 'lightning'

Hagege (1970: 161-162) also mentions more complex lexemes, some of which have phrasal structure, such as *pélé mì á nzì mbáp* 'tomorrow I ASPECT be.big too' = 'pinky.' Metaphor-driven conceptualizations in complex terms dominate in all domains, with the exception of artifact terms. In the artifact domain, like in (22c.), compounds with *fè* 'thing' abound. In fact, compounds with *fè* 'thing' and *nzùk* 'man, human' are the common way in Mbum to create agent and instrument nouns (Hagege 1970: 186), a function taken over by dedicated derivational morphemes in other languages.

4.5.1.2. Americas

4.5.1.2.1. *Miskito.* Languages with a similar lexical profile are also found in the Americas, although they are relatively rare in this area. One case in point is Miskito of the small Misumalpan family spoken in Nicaragua. This language features very little inflectional nominal morphology apart from possessive suffixes; the verb agrees with the subject and under certain conditions with the object in number, and the subject marking paradigm is intertwined with tense (Hale 1994: 264), as seen in (24.).

(24.) *Witin raks wal sula kum ik-an*
He gun with deer one kill-PAST3
'he killed a deer with the gun' (Hale 1994: 264)

In this language, morphologically complex items are exclusively of the lexical type; one such for an item of acculturation is seen in context in example (25.).

(25.) *Yang dusa pihni di-aia want sna*
I stick white smoke-INF want-PRES1[23]
'I want to smoke a cigarette.' (Hale 1994: 270)

Complex items in Miskito may involve verbal, nominal, and adjectival constituents (Conzemius 1929: 75 notes that "[m]any nouns, adjectives and verbs of the languages under consideration have the same root," which suggests that the distinction between noun and verb may be weak on the lexical level), as seen in examples in (26.).

(26.) a. *il byara* 'water abdomen' = 'valley'
b. *pauta yuya* 'fire grain' = 'spark'
c. *mita sirpi(ka)* 'hand little' = 'finger'
d. *ki mita* 'lock hand' = 'key'

[23] Additional gloss: PRES1 'present tense, 1st person.'

Terms for artifacts are either complex formations, as (26d.), or are loanwords from English, sometimes via Miskito Coast Creole English. Next to terms of the lexical type, Miskito also features derived terms formed with the suffix *-ka*, such as *aiklabaika* /aiklab-aia-ka/ 'fight-INF-DERIV' = 'weapon,' but they are clearly outnumbered by analyzable terms of the lexical type.

4.5.1.2.2. *Wappo.* The lexical type of complex lexemes is also dominant in Wappo, formerly spoken in California, and with this lexical profile the language is in contrast to many languages of North America. It would be incorrect to characterize Wappo as an isolating language since there is an apparatus of bound morphology associated with both verbs and nouns. As described by Thompson et al. (2005), nouns inflect for a number of cases, including the core cases nominative and accusative, while verbs are inflected by suffixes for tense/aspect and may carry directional prefixes. Mood is specified by preverbal particles. Wappo inflectional morphology is however comparably limited, lacking for instance the typical ingredients of polysynthesis such as incorporation and pronominal affixes cross-referencing information of NPs on the verb.[24] Pronouns are independent words rather than affixes, and there does not seem to be any overt agreement marking on the verb with its arguments. Furthermore, judging from the examples in Thompson et al. (2005), inflectional synthesis is rather low in actual usage, with inflection on nouns being mostly restricted to case, and most verbs in the examples bearing one or at most two inflectional morphemes. Wappo is included in this discussion, since its inflectional possibilities, while not non-existent, are limited when put in an areal-typological perspective, and there is a correspondingly higher degree of complex lexemes of the lexical type that appears to correlate with this situation.

Thompson et al. (2005: 9) mention associative phrases, consisting of two juxtaposed nouns (27.):

(27.) *oye? šukolo?-i* *pico:we-khi?*
pot bottom-NOM dirty-STAT[25]
'the bottom of the pot is dirty' (Thompson et al. 2005: 9)

This construction is also used to indicate inalienable possession, leading these construction to be indistinguishable from associative phrases (Thompson et al. 2005: 15), as seen in (28.).

(28.) *ċiċa khap-i* *ke?te-khi?*
bird wing-NOM broken-STAT
'The bird's wing is broken' (Thompson et al. 2005: 15)

[24] No rigorous attempt is made here to define the notion of "polysynthetic language" (see e.g. Mattissen 2003, Mithun 2009 for definitional issues).

[25] Additional glosses: NOM 'nominative case'

At any rate, combinations of two nouns, whether at the syntactic or at the morphological level, make up the largest proportion of analyzable lexical items in Wappo, although sometimes more complex structures are found (29c.) and the mechanism may be applied recursively (compare 29d. with 29e.):

(29.) a. *holchíla* /hól-chíla/ 'tree-skin' = 'bark, thin bark'
b. *huy-nán* 'breast-mouth' = 'nipple, teat'
c. *lékiš ċiti·-wélma* 'swallow bone-protection' = 'Adam's Apple'
d. *hu-méy* 'head-water' = 'tear'
e. *tuš-huméy* 'bee-tear' = 'honey'

4.5.1.3. Southeast Asia

4.5.1.3.1. *Vietnamese.* Languages of Southeast Asia are known for their isolating character, and indeed, the language used here to illustrate the prevalence of the lexical type of analyzable terms said to be associated with isolating languages, Vietnamese, is typically used in textbooks to illustrate a prototypically isolating language. However, obviously, in spite of the alleged one-to-one match between word and morpheme, Vietnamese is not devoid of analyzable items, particularly compounds. Thompson (1991: 126-138) provides an overview of different subtypes of compounds in Vietnamese, noting that in general it is quite difficult to distinguish them from syntactic phrases. The basic distinction is one in so-called syntactic and non-syntactic compounds (next to so-called "idiom compounds"): for the former, there are parallel syntactic phrases containing the same elements in the same order (one criterion available to distinguish them is differences in stress) and non-syntactic compounds, for which there is no parallel syntactic construction. Among syntactic compounds, there are those of the dvandva or coordinating type (called "generalizing compounds" by Thompson 1991: 128), such as *bàn-ghế* 'table-chair' = 'furniture' (paralleled syntactically by *bàn ghế* 'tables and chairs'), endocentric ("specializing") compounds, such as *dây thép* 'string/cord/wire-steel' (paralleled syntactically by *dây thép* 'electric wire). Among the nonsyntactic compounds is a type in which synonyms are concatenated (similar to Mandarin Chinese) called "reinforcing compounds;" of this type, only verbal compounds are given as examples by Thompson (1991: 130-131). Furthermore, among the nonsyntactic compounds there are endocentric compounds (called "attributive compounds" by Thompson 1991: 132), for instance *học trò* 'study school.age.child' = 'schoolchild, pupil'). According to Thompson, many examples of this type of compound are partially composed of loans from Chinese, and there are also some which are even calqued from Chinese. Chinese influence is also responsible for a class of compounds called "pseudo-compounds" by Thompson (1991: 133). This type involves bound morphemes (as defined by Thompson 1991: 118, that is, morphemes that cannot appear as a free form or in combination with a construction larger than a basic free form), some, but not all of which are borrowings from Chinese. Semantic relations in this type of compound are similar to those in non-pseudo-compounds (see Thompson 1991: 134-136).

As Thompson (1991: 136) notes, most of the compounds and pseudo-compounds are disyllabic (and hence, bimorphemic). Occasionally, however, compounding may apply recursively to yield words of three or four morphemes, such as *vô-tuyến điện-thoại* 'with-

out/lacking-wire/line electricity-speech/conversation' ='radio telephone.' However, a dispreference for such longer forms is noted, and shortening of three-syllable compounds to two syllables is observed (Thompson 1991: 137). This is in line with the data in the database. Combinations of two morphemes are clearly dominant in all semantic domains, although some words consisting of three elements can be found. Some examples are below (note that 30c. is trisyllabic/trimorphemic):

(30.) a. *lông chim* 'body.hair bird' = 'feather'
b. *sương mù* 'dew blind' = 'fog'
c. *dòng nước xoáy* 'current water swirl' = 'whirlpool'
d. *xe lửa* 'vehicle fire' = 'train'
e. *cổ chân* 'neck leg' = 'ankle'

There is also reduplication of various types (discussed under the heading "derivatives" by Thompson 1991: 139-140). Reduplication results in an iterative, augmentative, or other meaning, as in *nói nói* 'keep talking and talking' (*nói* 'talk').

4.5.1.3.2. *Yay.* Yay is another Southeast Asian language with the typical ingredients of the isolating profile. Grammatical relations are signalled not by affixes on either noun or verb, but by word order which is typically SVO:

(31.) *ku*[1] *mi*[4] *θoŋ*[1] *tua*[4] *ya*[5]
I have two CLF wive
'I have two wives' (Hudak 1991: xxvi, small caps added)

Example (31.) also demonstrates the absence of number marking on nouns, as well as the presence of a classifier system for most count nouns (Hudak 1991: xxvii). Verbs do not inflect for tense or number, and causatives are periphrastic:

(32.) *van*[4]*van*[4] *hauɥ*[3] *kwaat*[2] *hay*[6] *vaay*[4]
every.day cause sweep.out dung buffalo
'Every day he has him sweep out the buffalo dung'
(Hudak 1991: xxix, glossing slightly modified)

Another feature typical of isolating languages (though not exclusively found here) are serial verbs that may be thought of as making up for the absence of morphological means to express certain grammatical categories (Aikhenvald 2006: 53); Yay has them, as seen in (33.).

(33.) *may*[6] *faay*[4] *koŋ*[2] *ma*[1] *rup*[1] *caw*[3] *hauɥ*[3] *ku*[1]
bamboo bend come stroke head give I
'The bamboo bends down to stroke my head for me' (Hudak 1991: xxx)

The majority of analyzable items are made up of two elements with both nominal and verbal meanings. Since possession is expressed by the simple juxtaposition of possessor and possessed, such as *paw²po⁵ ti⁵* 'father he' = 'his father' (Hudak 1991: xxvi), there is no immediately obvious structural difference between these and possessive constructions, like in Wappo. Occasionally, however, concatenation of three elements is found.

(34.) a. *ram⁶ foŋ⁴* 'water roof' = 'wave'
b. *rua⁴ bin¹* 'boat fly' = 'airplane'
c. *ʔaay¹ ka¹ (raay⁴)* 'goiter leg (mark)' = 'calf'
d. *raan⁴ pay¹ rok⁵* 'house go outside' = 'outdoor toilet'

Quite often, analyzable terms in Yay are semantically redundant. This may be due to either one member of the complex lexical item already having the same semantic extension as the whole complex term, or the presence of a classifying morpheme.

(35.) a. *θɯay1 (ram⁶)* 'steam (water)' = 'steam'
b. *(θan³) rɯaŋ¹* '(CLASS.ROPE) tail' = 'tail'

Occasionally, there appear to be borderline cases in two senses: first, it is sometimes unclear whether a given morpheme is employed in a classifier-like fashion, and secondly, there may also be some differences in semantic nuances between simplex and complex terms. For instance, *ta⁵* 'river' may be combined with *ka¹* 'leg' to yield the meaning 'a river, a very big river.' If indeed *ka¹* 'leg' serves as a kind of classifier (which is suggested by the parallel construction *ka¹ ran¹* 'leg road' = 'road'), then the semantic difference between *ta⁵* and *ka¹ ta⁵* may be an instance of derivational usage of classifiers in Southeast Asia.

However this may be, it is worth pointing out that not all languages of Southeast Asia are exclusively of the lexical type. In another language, Sedang (Austro-Asiatic, Mon-Khmer), there is also a derivational apparatus exploitable for word-formation, including the nominalizing prefix *kơ-* and the nominalizing infix <*ơn*>. This feature is one of the main characteristics distinguishing many Mon-Khmer languages from other Mainland Southeast Asian language families which lack derivational morphology (Enfield 2005: 188), with some of the Mon-Khmer languages like Vietnamese shifting their typological allegiance due to areal pressure. In the database, derived terms in Sedang are virtually all restricted to the domain of artifacts; the vast majority of analyzable terms in other semantic domains are of the lexical type.

4.5.1.4. New Guinea

4.5.1.4.1. *Meyah.* In Meyah, of the East Bird's Head family, the inflection of nouns is very limited. Indeed, nouns "remain uninflected or have very little inflectional possibilities" (Gravelle 1998: 558). More precisely, animates may carry a plural suffix, and inalienable nouns obligatorily bear a prefix indicating person and number of the possessor (Gravelle 1998: 562; see Gravelle 2004: 114-130 for an overview of different types of nouns in Meyah). Verbs agree with either subject or object in person and number, and inflect for aspect,

mood and instrument (Gravelle 1998: 558). Both facts may be illustrated with the example in (36.).

(36.) *di-ra meiteb di-(e)r-agob ofa*
1-use machete 1S-INST-strike 3S
'I use a machete striking her/him'[26] (Gravelle 1998: 560)

Meyah lacks derivation as a major word-formation technique both in the verbal and nominal domain (Gravelle 1998: 564; for some minor instances of derivation see Gravelle 2004: 71). In the absence of derivation on a larger scale, verbs may be juxtaposed to specify or vary the meaning of the lexical items involved. For instance, the verb *eja* 'go' occurs in juxtaposition with adjective roots creating inchoative semantics (Gravelle 1998: 565), and similarly, in the absence of morphological causatives, *otonggum* 'make/do' and *agob* 'kill' are used to convey indirect and direct causation respectively (Gravelle 1998: 567).

Compounding operates by the juxtaposition of two lexical roots which may be of different parts of speech (Gravelle 1998: 569; see Gravelle 1998: 569-570 and Gravelle 2004: 130 for criteria for their identification). Abstract nouns in particular are formed by juxtaposition of *mar* 'thing' with other roots of different parts of speech (see also Gravelle 2004: 133):

(37.) a. *mar ɸ-oga ɸ-en-ah-ma*
thing 3S-word 3S-DUR-sits-DEM
'There is a word thing' = 'There is a dispute'

b. *ofa ɸ-odou ɸ-ebriyi gij mar moisu*
3S 3S-liver 3S-splits in thing shaman
'His liver splits in shaman things' = 'He knows about magic'
(adapted from Gravelle 1998: 570)

Example (37b.) also illustrates the frequent use of body-part terms, in particular *odou* 'liver,' for the expression of psychological states.

The formation of noun-noun compounds involving two inanimate nouns or one inanimate noun and a noun denoting a part of the body is also "highly productive" according to Gravelle (1998: 570). It is this type which makes up, in the absence of any derivational processes, the entire set of morphologically complex lexical items in all semantic domains in the database. However, N-V compounds are also attested (Gravelle 2004: 132). While complex terms consisting of two roots are most frequent, there are some items in which as many as three are combined (Gravelle 1998: 564). As noted in general by Gravelle (1998: 130), "Meyah productively uses compounding to express nominal meanings that can be glossed with one word in another language, such as English." Examples include:

[26] Additional gloss: INST 'instrument'

(38.) a. *mocgój efésa* 'fog/cloud flash' = 'lightning'
b. *méngk ofód* 'breast flood' = 'milk'
c. *mówa eitéij* 'sun eye' = 'clock'
d. *ojóna otkonú* 'married.woman stomach' = 'womb'
e. *mówa ot déis* 'sun stand middle' = 'noon'

Compounds of either type function syntactically like simplex nouns in that they may occupy argument positions, as in (39.).

(39.) *Ri-of terapan nou kabar ofu ke-uma*
3PL-cover tarp for ship fly NOM-that[27]
'They covered that plane with a tarp' (Gravelle 2004: 132)

While Meyah does feature systems of numeral and noun classification (Gravelle 2004: 137-142), these systems do not appear to be employed in a derivational fashion frequently.

4.5.1.4.2. *Toaripi.* A language with a very similar lexical profile in New Guinea is Toaripi, belonging to the small Eleman language family. Here, too, morphologically complex terms consisting of two lexical roots dominate, as seen in (40.).

(40.) a. *fave uta* 'stone/rock hole' = 'cave'
b. *ori tolo* 'bird leaf' = 'long wing- or tailfeather'
c. *a-e* 'fire-faeces' = 'spark'

Occasionally, however, one encounters complex lexical items in which the two constituents are linked by a possessive marker, such as *ofae ve mehe* 'eye POSS hair' = 'eyelash.' Also occasionally, constituents appear in reduplicated form in the morphologically complex lexical item, such as in *maea ma ma* 'sweat, perspiration' (*maea* 'body,' *ma* 'water'). Like in Meyah, artifact terms involving 'thing' are not encountered for the items on the wordlist (but given that they do exist in Meyah, they may also exist for other meanings in Toaripi); artifact terms in Toaripi are either other morphologically complex terms of the lexical kind, or are designated by semantic extension (such as *ma ove* 'water picture,' originally 'reflection in water' and then extended to 'mirror, looking-glass'), or loanwords from Motu or English.

4.5.2. COMPOSITIONAL TYPE OF POLYSYNTHESIS: KET

Vajda (2004a: 400) notes a descriptive gap in traditional morphological typology in that there is no standard term for languages with a rich inflectional, but relatively sparse derivational apparatus, in contrast to languages such as those of the Eskimoan and Wakashan families, which also have a large amount of derivational morphemes. These are prime examples of Mattissen's (2003, 2004) affixal type of polysynthesis, and are here representa-

[27] Additional Gloss: NOM 'nominalizer.'

tives of the derived type with respect to their lexicon, a fact which corresponds well with the wealth of available derivational mechanisms. Vajda suggests "conglomerating language" to describe the other language type he has in mind. His example is the Yeniseian language Ket, which is used as an example of the compositional type of polysynthesis in Mattissen (2004), meaning roughly that it allows the concatenation of more than one lexical root into the noun complex. Notably, the difference noted by both authors with regard to the positioning of Ket on the spectrum of polysynthesis when compared with languages of the Eskimoan and Wakashan families corresponds to a difference in the lexical profile for Ket in that the overwhelming majority of complex nominals in the database are of the lexical rather than of the derived type.

According to Vajda (2004a: 413), Ket favors conversion as a transcategorical, word-class changing operation. However, a nominalizing suffix *-s* does exist: examples are *nánbèt-s* 'baker' (*nánbèt*, 'to bake bread') and *úgdè-s* 'length' (*úgdè* 'long,' see also Vajda 2004b: 15 and Werner 1998: 39-41 for further discussion and examples). Vajda (2004b: 15) also mentions some archaic non-productive noun derivation devices as well as quasi-derivational use of terms for 'male' and 'female.' The nominalizing suffix *-s* is responsible for the few terms in the database classified as being of the derived type in the database, of which an exhaustive list can be found in (41.).

(41.) a. *də-doq-s* '3N-fly-NMLZ' = 'airplane, helicopter'[28]
b. *binaʁols* /bin-o-qol-s/ 'self-PST-heal-NMLZ' = 'scar'

In addition, there is redundant marking of plural on some nouns (Vajda 2004b: 20; according to Vajda 2004b: 80, "multiple plural suffixes in many nouns came to serve as stem-building elements," that is, are employed in a fashion remotely resembling derivation, see Werner 1998: 41-43 for further discussion). This phenomenon occurs for instance in *ulteɣin* /ul-te-in/ 'water-lake-PL' = 'swamp, quagmire;' a further case seems to be *cas-aŋ* 'hour/watch-PL' = 'clock, watch, timepiece.'[29]

However, compounding is a much more frequent word-formation strategy in Ket, and indeed Werner (1997: 46) states that compounding must be an old mechanism in the language since most polysyllabic and even some monosyllabic words can be traced back historically to compounds. Werner (1998: 49-50) distinguishes between proper and improper compounds ("eigentliche und uneigentliche Komposita"). For the former type, bare roots are concatenated, while for the latter type (which is particularly common if the relationship between the roots is possessive or synecdochic according to Vajda 2004b: 15), the genitive marker *-d* occurs as a linking element. Werner (1998: 50) notes that this type of construction is not typical for Yeniseian languages. Examples for compounds without such a linking element are seen in (42.).

[28] Ket also has the borrowed terms *samolop* and *eroplan* (from Russian *samolët* and *aeroplan* respectively).
[29] As Bernard Comrie (p.c.) points out, it may be relevant that in Russian *časy* 'watch, clock' is grammatically the plural of *čas* 'hour.'

(42.) a. *mamul* /maˀm-ūl/ ‘breast-water’ = ‘milk’
b. *espul* /ēs-hɯl/ ‘sky-mound’ = ‘cloud’
c. *aqqot* /ekŋ-qoˀt/ ‘thunder-path’ = ‘rainbow’
d. *eŋgaj* /eˀŋ-àj/ ‘eggs-sack’ = ‘nest’

Werner (1998: 53) notes that elements such as *-ūl* ‘water’ (as in 40a.) and *-àj* ‘sack’ in example (40d.) (as well as a number of others) may be considered to be halfway along a grammaticalization path to derivational elements, and calls them word-forming semi-affixes (“wortbildendes Halbaffix”). The reasons for such a position are that they occur in a series of compounds while at the same time diverging phonologically from their free-standing counterparts (a point also noted by Vajda 2004b: 14). However, Werner (1998: 58) states that since they can be readily identified with lexical sources, they should still be regarded as elements of a compound rather than derivational suffixes. Examples of compounds with the linking element *-d* include:

(43.) a. *bulaŋdoks* /būl-aŋ-d-ōks/ ‘leg-PL-POSS-wood’ = ‘ladder’
b. *destqār* /dēs-d-qār/ ‘eye-POSS-hair’ = ‘eyelash’
c. *sestbaˀŋ* /ses-d-baˀŋ/ ‘river-POSS-land’ = ‘coast’
d. *oləndquk* /olən-d qūk/ ‘nose-POSS hole’ = ‘nostril’

Vajda (2004b: 15) also discusses a number of examples of compounds which are unlike the examples above complex words on phonological grounds, but rather phonological phrases.

Thus, Ket, which is typologically quite different from neighboring Eurasian languages, also stands out when compared to the typically mixed lexical-derived profile of its neighbors (see § 4.6.6.), as well as by the presence of some highly complex nominals that go well beyond the combination of two roots, such as *uldʌptaŋbaŋ* /ul-də-b-taŋ-baŋ/ ‘water-3N-3N-turn-place’ = ‘whirlpool.’ Furthermore, the Ket lexicon, at least judging by the equivalents to the meanings on the wordlist used here, is characterized by an elevated degree of analyzable terms in the broader Eurasian context (in spite of Vajda’s 2004b: 14 comment that “Ket basic vocabulary includes numerous non-derived stems, many of them monosyllabic”). To conclude the discussion of Ket, the following textual example (from Vajda 2004b: 97, boldface added) illustrates a compound of the type with the genitival linking element in the context of discourse; in addition this example gives an idea of the polysynthetic nature of the Ket verb, with the morphological analysis provided in square brackets and position classes indicated by superscript numbers:

(44.) *bū* *ttíŋgòmdaq* *[du^{8}-tiŋ^{7}-k^{5}-o^{4}-b^{3}-in^{2}-daq^{0}]*
3M.SJ he.stopped.those.up [3M.SJ^{8}-plug^{7}-ADES^{5}-D^{4}-3N.O^{3}-PT^{2}-shove^{0}]

ít-dès ***d-ólàn-d*** ***qúk-sèn*** *haj* *ít-qò*
3F.POS-eyes **3F.POS-nose-N.GEN** **hole-PL** and 3F.POS-mouth
'He stopped up her eyes, **nostrils**, and mouth with sand.'[30]

Nivkh has a similar lexical profile, although there are differences in the structure of the noun complex (Mattissen 2004: 51, see the extensive discussion of Nivkh in Mattissen 2003). Also similar is the lexical profile of Cheyenne, although here a certain amount of deverbal derivation exists (see also § 4.6.2).

4.6. LANGUAGES WITH MIXED PROFILES

4.6.1. OVERVIEW

As stated in the introduction to this chapter, languages relying exclusively on complex lexical items of either the lexical or of the derived type only represent the extreme end-points of a typological continuum, and many languages (in fact, most languages) in the sample make use of both types in the creation of morphologically complex expressions and thus fall somewhere in between the extremes. This section aims to highlight some different types of languages with such mixed profiles, but it cannot do full justice to the actual diversity found. Another complication is that many, but not all, languages of this type have a very low proportion of analyzable lexical items, and thus their classification is due to a rather restricted number of examples. For instance, in Imbabura Quechua, there is an absolute number of two lexical items of the derived type in the database, which amounts to a relative number of 8.9 per cent of derived-type lexemes among the analyzable lexemes of that language. Note that in absolute numbers this is precisely the same figure as found in Ket, but due to the higher total number of analyzable terms (compounds, in this case) in Ket, it falls into the category for which the lexical type is dominant, and not in the mixed-type category. The fact that the languages in this group are so typologically heterogeneous is not least due to the comparatively low total number of analyzable terms. To illustrate this, a discussion of languages on different ends of the typological continuum (but not at the extremes) on the syntheticity-analyticity scale, Efik and Hawaiian on the one hand and Sora on the other, follows.

[30] Glosses: ADES 'adessive case suffix; in finite verbs, a derivational affix denoting motion towards,' D 'durative marker (appears in many stative and activity verbs),' F 'feminine class (a subset of animate class),' POS 'possessive clitic (proclitic on noun phrases; derived from genitive-case suffixes,' M 'masculine class,' N 'neuter (=inanimate class); either singular or plural,' O 'verb-internal direct-object agreement affix, or direct-object pronoun,' PL 'plural,' PT 'past tense,' SJ 'verb-internal subject agreement prefix, or subject pronoun.'

4.6.2. MORE OR LESS ISOLATING PROFILE: EFIK AND HAWAIIAN

A case of a mixed-type language with a relatively low degree of inflectional synthesis is Efik, a Benue-Congo language of Africa. Efik has complex expressions of the lexical type, as seen in (45.).

(45.) a. *akam'ba obüt'* 'great land/earth' = 'mountain'
b. *idib'i uküt* 'belly/womb leg' = 'calf'
c. *ök'pö ñkañ* 'bone side' = 'rib'

However, Efik seems to prefer to combine only nouns in this way. When complex items involve verbal roots, a nominalizer in the form of a prefixed vowel or syllabic nasal exhibiting a wide range of allomorphy is typically employed (Welmers 1968: 157-158, Essien 1990: 119-121), as in (46a. - 46d.). There is also the prefix *eri-,* which forms participial nouns or gerunds from verbs (46e.)

(46.) a. *e-fak'* 'NMLZ-fix.between' = 'a narrow street, a lane'
b. *e-kep'-kep* 'NMLZ-red-to.flash' = 'flashing, lightning, glistering'
c. *u-tök' ikïm* 'NMLZ-to.void urine' = 'bladder, chamber pot'
d. *ukur'isü* /u-kut-i'sü/ 'NMLZ-see-face' = 'looking glass, mirror'
e. *eritem'* 'ABSTR-cut.down' = 'clearing'

As can be seen from examples (46c.) and (46d.), these deverbal terms sometimes are accompanied by a nominal element which may be conceived of as the object of the nominalized verb (Welmers 1968: 157). The deverbal formations in general are subject to lexicalization to a large degree; it is frequently the case that the initial vowel of a noun is analyzable as a nominalizer, but the following stem does not occur outside of the construction (Welmers 1968: 157, Essien 1990: 116 on the closely related language Ibibio).

Languages with a similar profile include other African languages such as Hausa, as well as Austronesian languages such as Tetun and Hawaiian. In addition to partial and full reduplication, Hawaiian employs a variety of derivational morphemes to derive meanings from morphologically unanalyzable terms (47.). However, as seen in example (47b.), derivational suffixation may also be applied to a reduplicated form. Examples include:

(47.) a. *ʻoi-ʻoi* 'RED-sharp' = 'thorn'[31]
b. *ʻoiʻoi-na* 'pointed/protrude-NMLZ' = 'headland'
c. *kā-maʻa* 'CAUS-bind' = 'shoe'
d. *mō-lehu* 'QUAL/STAT-ashes/ash-coloured' = 'dusk'

It is perhaps noteworthy that in languages with a mixed profile and a relatively low degree of synthesis, in particular in the domain of nominals, one of the derivational processes employed is or appears to be frequently full or partial reduplication. Examples from a variety of different sample languages are:

[31] Both *ʻoiʻoi* and *ʻoi* also have other related meanings.

(48.) a. Hawaiian *lihi-lihi* 'eyelash, eyelid' (*lihi* 'edge')
b. Sahu *so-soloro* 'bird' (*soloro* 'to fly')
c. Bororo *baru-baru* 'horizon; type of spirit' (*baru* 'sky')
d. Rama *kung-kung* 'lungs' (*kung* 'louse, whit, air root')
e. Bislama *kol-kol* 'fog, mist' (*kol* 'cold')[32]

However, except for Sahu and Hawaiian, reduplication is not the dominant derivational strategy in any of the languages in this survey, and therefore there is at most a weak correlation between mixed-type languages with reduplication as a derivational device.

4.6.3. COMPOSITIONAL TYPE OF POLYSYNTHESIS: SORA

The Munda language Sora has certain typological similarities with Ket in that it belongs to the compositional type of polysynthetic languages. Details differ, however. Sora verb morphology is extensive, and Sora nouns inflect for number, person, and possession, and their function in the clause is indicated via case and case-like adpositions (Anderson and Harrison 2008: 306). In terms of word-formation, nominal roots in their combining form play a major role. Together with a large number of derivational processes including affixation and reduplication, concatenation of such nominal roots (up to three or sometimes even four roots according to Anderson and Harrison 2008: 307) serves to create the nominal lexicon of Sora, and more than one derivational element may occur in the same form (Anderson and Harrison 2008: 307, 327; see also Anderson and Harrison 2008: 330 for a "monster noun form" featuring four derivational infixes). Combining forms may also be incorporated into the verb complex (most often, the incorporate corresponds to the patient of a transitive verb, Anderson and Harrison 2008: 355), and even here, more than one combining form may be concatenated, as example (49.), from Anderson and Harrison (2008: 359) who quote Ramamurti (1931: 143), shows:

(49.) *jo-me-bo:b-dem-te-n-ai*
smear-oil-head-RFLXV-NPST-ITR-1
'I will anoint myself with oil'[33]

Examples of complex nouns of a variety of types in Sora are seen in (50.). These include concatenation of combining forms (50a. - 50d.), sometimes including additional material such as a reflexive morpheme in (50b.).

[32] Both *kolkol* and *kol* also have other meanings.
[33] Additional glosses: ITR 'intransitive,' NPST 'non-past,' RFLXV 'reflexive.'

(50.) a. *əsu:ŋtidən* /ə-'su:ŋ-'tid-ən/ 'POSS-hut.for.temporary.use-bird-N.SFX' = 'nest'[34]
b. *duŋdəm'da:(ba:)n* /duŋ-dəm-dˀa:-(ba:)-n
'get.out.of-RFLXV-water-(place)-N.SFX' = 'a natural spring of water'
c. *'kuru:tamən* /'kuru·-'tam-ən/ 'body.hair-mouth-N.SFX' = 'beard'
d. *dai'jʊŋən* /dai-jʊŋ-ən/ 'climb-sun-N.SFX' = 'sunrise'

4.6.4. DERIVED TERMS IN THE DOMAIN OF ARTIFACTS AND STRUCTURAL HOMO- AND HETEROGENEITY IN DERIVATIONAL DEVICES

In spite of the large structural heterogeneity of languages of the mixed type noted above, it is still possible to discern certain patterns in their behavior. In some languages of this type, derived terms for the meanings under investigation are very common or even found exclusively in the domain of artifacts, but are rare or unattested in the other domains. This is true for instance in Chickasaw, in which two thirds of derived terms cluster together in the domain of artifacts. The two most common basic schemes are locative nominalization with the prefix *aa-* and the nominalizing suffix *-'*, and usage of *isht* 'with' together with the same nominalizing suffix *-'* for instrument nominalization. As the example in (51b.) makes clear, the basic types may be amended by additional bound morphological material:

(51.) a. *aa-nosi-'* 'LOC-sleep-NMLZ' = 'bed'
b. *aa-ili-pisa-'* 'LOC-REFLX-see-NMLZ' = 'mirror'
c. *ishtalakchi'* /isht talakchi-'/ 'with be.tied.up-NMLZ' = 'rope'

Santiago Mexquititlan Otomí is a mirror image of this in that derived terms are not found in the domain of artifacts at all (the language has largely borrowed its vocabulary in this domain from Spanish), but are dominant in the domain of nature-related lexical items:

(52.) a. *mbo-zaa* 'place.where.around-tree' = 'forest'
b. *ma-hets'i* 'LOC-high' = 'sky'
c. *h-ñä* 'IMPERSONAL.VOICE-speak' = 'steam, voice, language'
d. *munts'i* /m-punts'i/ 'NMLZ-to.relapse' = 'whirlpool'

Similarly, in Wichí, analyzable terms in the domain of nature-related terms are almost exclusively of the lexical type (53a., 53b.), with the exception of the term for 'rain,' which consists of a subject marker prefixed to the verb root *-wu-* 'to make' and followed by an epenthetic morpheme and a locative suffix (53c.).

(53.) a. *itoj muwk* /itoj mukw\n/ 'fire dust' = 'ashes'
b. *tewuk lhip* 'river part' = 'coast'
c. *i-wu-m-cho'* '3SBJ-make-EP-LOC.under' = 'rain, to rain'

[34] Gloss in the source is 'a nest of binds' [sic].

In contrast, Wichí terms for artifacts and body-parts are often of a different structure; some of them involve prefixation of the possessive marker *to- ~ tot-*. The terms of the derived type in this language occur frequently in combination with this possessive marker, as seen in the examples in (54a., 54b.), (although there are also complex terms of the lexical type with this prefix as well as artifact and body-part terms which do not feature it, like for instance 54c.).

(54.) a. *tot-telhu-hi-s* 'POSS.INDET-eyes-LOC.in-PL' = 'glasses'
b. *to-nhes-pe'* 'POSS.INDET-nose-LOC.on' = 'nostril'
c. *y'amekw-hi* 'excrement-LOC.in' = 'toilet'

San Mateo del Mar Huave is a Mesoamerican language that behaves similarly.

Furthermore, another independent variable is that, in some of the sampled languages, derived terms are mostly construed by use of one and the same derivational marker. This is the case in Chickasaw, for which the majority of such terms involve the nominalizer *-'* (often in conjunction with *aa-* 'locative' or *isht* 'with'), while in other languages, such as Santiago Mexquititlan Otomí, terms classified as being of the derived type are of a variety of structural types.

Table 2 cross-classifies languages with mixed profiles according to these two variables. The criterion for assignment is whether 50% of the derived terms are in the domain of artifacts or not, and whether 50% of them are formed by the same derivational device or not. This is not always possible, for instance when the number of derived terms is equally distributed over the artifact domain and other domains or when the number of derived terms is precisely one. Languages in which one of these situations obtains are not shown in the table. For languages in which one derivational device dominates, the corresponding morphemes and glosses are given in parentheses.

	different devices	mostly same device (given in brackets)
mostly artifacts	4 Nuuchahnulth, Carrier, Basque, Bororo	4 Hausa (*ma-...-i* 'INSTR-...-INSTR'), Chickasaw (*-'* '-NMLZ'), Haida (*-ʔu* '-INSTR'), Guaraní (*-ha* '-AGT'),
mostly or exclusively non-artifacts	16 Great Andamanese, Pipil, Arabela, Kiliwa, Kiowa, Hawaiian, Welsh, Kanuri, Pawnee, Jarawara, Cavineña, Santiago Mexquititlan Otomí, Itzaj, Imbabura Quechua, Embera, Bislama	4 Sahu (reduplication), Sora (*-ən* '-N.SFX'), Kolyma Yukaghir (*-i:* '-INSTR'), San Mateo del Mar Huave (*-aran* '-INAL.POSS')

table 2: languages with a mixture of complex terms of the derived and lexical type and their structural diversity and spread over semantic domains

As is clear from the table, there is a strong skewing for derived terms to not be in the domain of artifacts in languages with many different devices to form complex words of the derived type. Specifically, there are only four languages in which these are found domi-

nantly in the artifact domain; in contrast, in the vast majority of such languages, the terms are more frequent in other semantic domains. Among them are Pipil, Santiago Mexquititan Otomí, and Itzaj, which have borrowed terms for artifacts from contact languages, in most cases Spanish, rather than coining neologisms (§ 5.4.2.7.1. deals with borrowing vs. coinage of neologisms, the influence of the contact language as well as structural factors in more detail), but also Kiliwa and Kiowa, which did the precise opposite. At any rate, there is a skewing with respect to languages with high structural diversity in derived terms when cross-classified with the semantic domains these are found in that is not observable in languages where one structural type dominates.

Apart from this classification, there are also commonalities shared between some languages with a mixed profile. Whether these stand in a direct relationship to their assignment to the mixed type is a question that must remain open for the time being; nonetheless, these commonalities by themselves are striking enough and of interest for lexical typology that they are worth discussing. The first commonality is areally and perhaps, on a more subtle level, even structurally motivated, and pertains to complex nominals with a clause-like structure in many sampled languages of North America (§ 4.6.5.). The other is areal only and stems from the observation that many languages of Eurasia rely on a mixture of derived and lexical terms in their morphologically complex terms (§ 4.6.6.).

4.6.5. Verb-centered Nominals in North America (and beyond)

4.6.5.1. Overview

Mithun (1999: 287) notes for North American languages in general in the context of the discussion on ceremonial speech that "[t]he structure[s] of many North American languages ... lend themselves well to such periphrastic expression, due to their general polysynthesis. Single words are often composed of many meaningful parts, and their literal meanings are in many cases still perceptible to speakers. Indeed, words in all domains are frequently coined from complex descriptions ..." While complex nominals in North America have a wide range of structural types, as seen above, there is one particular type which is quite common in languages spoken in North America, although the phenomenon in question is neither found in all languages of the continent nor does it need to be the primary type for the formation of complex nominals. The type in question are clausal nominals that are formally either similar to headless or internally headed relative clauses or are fully inflected verb forms which may sometimes, given the appropriate context, receive a compositional interpretation as such, but which have specialized institutionalized semantics to denote a particular object or entity. Languages in which either of these are found typically fall into the mixed category due to the mixture of verb-based derivation involving a single root and the clausal structure involved in other terms involving more than two roots. Such languages have no clear lexical profile. In spite of this, the similarity of these devices is evident enough to justify the postulation of a separate type of word-building, due also to it appearing as either areally or typologically restricted.

4.6.5.2. Clausal Nominals with Relative Clause-like structure

4.6.5.2.1. *Kiliwa.* Kiliwa, a Yuman language of Baja California, is a good example for illustrating nominals with relative clause-like structure, as well as for showing that this strate-

gy typically coexists with other types of word-formation, which together account for the morphologically complex expressions in the lexicon. Kiliwa nominals, for instance, may be compounds that can either have nouns or adjectives as constituents:

(55.) a. *nmi ʔ=tay* 'cat=large' = 'cougar; puma'
b. *haʔ=ñmiʔ* 'mouth=fur' = 'moustache; beard' (Mixco 2000: 25)

More importantly, however, there are mechanisms that allow for the formation of deverbal nouns, of which Kiliwa has a wide variety. As discussed by Mixco (1965: 100), there are traces of archaic and non-productive types of nominalization that highlight the essentially verb-oriented character of the language, such as a vowel shortening rule evidenced by some striking resemblances of some nouns to verb roots, e.g. *n-míʔ* 'cat' (cf. *(ʔ)-mîˑ* 'to cry') and *ʔ-wáʔ* 'house' (cf. *wâˑ* 'to dwell). Further evidence for a verbal origin of a large number of Kiliwa nouns is provided by nouns which seem to be analyzable into a verbal root and a verb prefix, thus not showing any sort of overt marking for nominalization, such as *č-pàt* 'doorway' (*č-* 'perpendicular axis prefix found on verbs,' *pàˑ* 'to leave'), *m-phú* 'hole' (*m-* 'stative,' *-phú* 'to perforate,' Mixco 1965: 93-94). Furthermore, there is evidence for lexicalization of erstwhile nominal compounds with a phonological contraction in the first element. Mixco (1965: 106) derives the first element *w-* of *w-xâˑ* 'cave' and *w-ñáʔ* 'road' diachronically from **wîˑy* 'mountain, stone(?).' In fact, Mixco (1965: 92) characterizes the entire nominal lexicon of Kiliwa by saying that "[n]ouns either give the impression of being eroded nominal compounds or atrophied verbs, with fossilized affixation and traces of vestigial verbal processes such as reduplication and vocalic ablaut."

Synchronically analyzable morphemes to derive nouns from verbs include:

(i) *t-* 'reifier,' as in *t-híp* 'grass, fodder' (*-híp* 'to sprout') (Mixco 1965: 95-96; glossed OBJ in Mixco 1985)
(ii) *-uʔ* 'locative/nominalizer.' Terms derived with this suffix frequently have a locative meaning, as for instance in *h-pâˑ-uʔ* 'sleeping area' *(-pâˑ* 'to lie down'), but it is also employed as a general nominalizer (Mixco 1965: 96). Derivations with this suffix may in turn enter into larger structures such as nominal compounds, such as *mñis w-uʔ* 'foetus sit-LOC' 'womb' (a term taken from the database)
(iii) *-tay*. This postclitic is identical to the verb *-tay* 'to be large' and has a nominalizing function as in *hphíʔ táy* 'musical instrument' (*phíʔ* 'to make sound'), with the resulting noun denoting someone or something that performs the action of the verb it is derived from frequently (Mixco 1965: 97).

However, the most productive morpheme in the language in general according to Mixco (1965: 95) is the prefix *k*^{w}-, which derives nouns from adjectives with the meaning 'one characterized by x' or 'one that does x,' where x is the meaning of the verb root, as in the examples in (56.).

.

(56.) a. *phi? k^{w}-msúx* 'nostril' = 'nose which is perforated' (Mixco 1965: 102)
b. *k^{w}-phú·y* 'smoke, grey' (*-phû·y* 'be smoky') (Mixco 1965: 95)

Nominalization with *k^{w}-* is also the most interesting process from a typological point of view since *k^{w}-* also functions as a relativizer. In the following more complex examples (from Mixco 2000: 25), the overlap in function between nominalization and relativization becomes particularly clear. All examples involve a further nominal root which is modified by a nominalized verb by means of *k^{w}-*, or, since the process of nominalization and relativization appears to be indistinct in these cases, a relative clause introduced by *k^{w}-*.[35] Mixco (1965: 101) calls such structures "relativized compounds."

(57.) a. *xaq-m=k^{w}-sk^{w}aayu* 'beef-OBJ=WH-guard' = 'cowherd'
b. *myal=k^{w}-sn?aawu* 'tortilla=WH-small:PL' = 'cracker; cookie'
c. *?wa?=k^{w}-s?hin* 'house=WH-run' = 'automobile'
d. *xwa=?iy=kw-kuus* 'enemy=hair=WH-long:PL' = 'Cocopa'

But there is more to be said about deverbal derivation in Kiliwa. Notably, two or more nominalizers may be combined. Mixco (1965: 97) calls this characteristic "complex nominalization." Examples are given in (58.), with the derivational morphemes highlighted by boldface (see also Mixco 1965: 97-98 for examples 58a. and 58b.)

(58.) a. ***t-kw-****ipaa-y* 'OBJ-WH+be.alive-ATT' = 'animal'
b. ***t-****xpapu-****u?*** 'OBJ- sew-OBL' = 'needle'
c. *wa?=t-****kw-****lkwii-y-****tay*** 'house-OBJ-WH-carry-ATT-FREQ' = 'train'

Furthermore, there is a type of relativized compound in which there is no overt mark of nominalization according to Mixco (1965: 101). Examples provided by Mixco (1965: 102) only include terms for animals, such as *?múw há? ñmí? k^{w}yúw* 'sheep (which has) (on the) mouth hair which stands' = 'goat,' so it is unclear just how productive this type of nominalization is in Kiliwa.

Kiliwa compounds readily combine with bound nominal morphology, such as the demonstrative and illative suffixes *-mi* and *-l* in example (59a.). In this example, the compound seems to be in apposition to the actual clause syntactically. However, Kiliwa compounds also readily occupy argument positions, as seen in (59b.), where the compound is marked as subject by the suffix *-t*:

(59). a. *mi=?-m-mi-l* *č?ii-k,* ***miy=ha?****-mi-l*
this=DN-COM-this-IL stick-DIR, **leg=face**-this-IL
'An arrow stuck right here in his **calf** [Achilles spot].'
(Mixco 2000: 58, boldface added)

[35] According to Comrie and Thompson (2007: 379), indistinctiveness of nominalization and relativization is found in a number of languages in Western and Southwestern North America (see further references there).

b. *yu-m=yuu-t, h-paa* ***yuw=xaʔ****-t* *xwaq-uʔ-l* *xiʔwap-m-t*
be-DS=be-SS, 3-recline **eye=water**-SBJ two-REL-IL spill-DIR-SS
'And so, as he lay weeping, his **tears** spilled on either side
(of the watershed)'[36] (Mixco 2000: 52, boldface added)

It would be very interesting to know how other types of complex nominals, in particular those involving *kʷ-* behave syntactically, but unfortunately, no such example occurs in the short text samples provided in Mixco (2000).

4.6.5.2.2. *Cheyenne.* In Cheyenne, the marker *tsé-*, glossed in the source as 'that which is,' plays an important role, in particular in complex lexemes in the domain of nature-related terms. It may be prefixed to a verb root to form the name for a topological feature, as in (60a.). Its apparent relativizing function becomes clear by the fact that verbs prefixed with *tsé-* may be accompanied by a simplex noun that it modifies, with the derived verb either following (60b.) or preceding the noun (60c.).

(60.) a. *tsévé'evótoo'e* /tsé-vé'evótoo'e/
'that.which.is-be.a.concave.hole' = 'cave'
b. *tsésééha mahpe* /tsé-séeha mâhpeve/
'that.which.is-spread.out water' = 'lake'
c. *hotohke tséana'ôhtse* /hotohke tsé-ana'ôhtse/
'star that.which.is-fall' = 'meteoroid'

However, in the domain of artifacts, unlike in some other languages, the construction apparently plays a minor role. Here complex formations with *hestôtse* 'thing' dominate:

(61.) a. *táxemésêhestôtse* /táxe-mésehe-hestôtse/
'upon-eat-thing' = 'table'
b. *ameohestôtse* /ameohe-hestôtse/
'go.by.quickly-thing' = 'car'

In addition, Cheyenne also features a large number of compounds, such as *háhnoma-pano'êhasëö'o* /háhnomâ-hépano'êhasëö'o/ 'bee-syrup' = 'honey' (compounds in fact outnumber the deverbal type of complex lexemes), leading to the coexistence of many different structural types of morphologically complex expressions in the lexicon.

4.6.5.2.3. *Carrier.* In Carrier, a candidate for a morpheme which both has a relativizing function and is used to coin complex nominals is *-î,* glossed here as 'REL' in the following instances.

[36] Glosses: COM 'commitative [sic!] (with),' DS 'different subject,' IL 'illative (in; inside),' REL 'non-subject relativizer,' SS 'same subject.'

(62.) a. *hananelyih-î* 'grow.again-REL' = 'seed'[37]
b. *dzeł-î-krez* 'mountains-REL-between' = 'valley'

However, there appear to be other structural types of verb-based nominals. Morice (1932: 84-85) points to terms such as *uḳwetsežthi* 'bed,' literally translated as 'on it one lies down' and *uṭaz teratsiyaih* 'ladder,' literally 'by the help of it one goes up.' These are, in the words of Morice (1932: 84), "nothing but a verb preceded by a pronominal preposition without a substantive," and it is these examples which provide the transition to another structural type of complex nouns found in some languages of the world, in particular those of North America: derived nouns without any overt indication of the derivational nominalizing process.

4.6.5.3. Nominalization without Nominalizing Morphology

One phenomenon encountered frequently in languages of North America concerns expressions with nominal meanings which are morphosyntactically similar or identical to finite verb forms. This type of construction is also the structural phenomenon that underlies personal names such as 'dances with wolves,' and it will thus be referred to as the 'dances with wolves'-type of nominal mnemonically. It is "a commonplace for Americanists" according to Beck (2005: 3), and indeed, such constructions are found in a number of sampled languages.

4.6.5.3.1. *Kashaya.* In Kashaya, derived nominals with the same inflection as found on verbs may be formed by a morpheme called the absolutive (this is an entirely different type of marker than the eponymous absolutive marking S and P in languages with ergative alignment, and should not be confused with it). This suffix, which has a wide range of phonologically conditioned surface allomorphy, is the citation form for verbs. In narratives, the absolutive form of the verb is the most frequent form of the main verb (Oswalt 1961: 266). In other words, a verb inflected with the absolutive is clearly finite. However, verbs inflected with the absolutive also have a nominal function, and "[t]he absolutive behaves syntactically just like a common noun" (Oswalt 1961: 266). For instance, as seen in example (63.) from the database, it may enter into larger constructions with other nouns.

(63.) *q^{h}a·moš šuṭ^{h}uhṭadu* /q^{h}a·moš šu-hṭ^{h}uṭ-ci^{O}d-w/
'star by.pulling-pieces.come.off.bigger.object-DUR-ABS' = 'shooting star'

This is the only example in the database in which the absolutive morpheme is found in terms from the domain of nature and topology. It is more frequent in terms for artifacts, particularly when combined with a following element *=li* (here tentatively analyzed as a postclitic, but possibly an affix or a postposition), indicating instrument or location.

[37] The "literal" translation provided by the lexicographer for *hananelyih-î* is "that which uses to [sic!] grow again." Presumably "uses to" is meant to highlight the habituality; however, the present discussion is only concerned with the fact that the literal translation employs a relative clause.

(64.) a. *caduli* /cad-w=li/ 'see-ABS=INSTR' = 'mirror'
b. *caq^{h}awli* /ca-q^{h}a-w=li/ 'WITH.MASSIVE.OBJECT-cut.off-ABS=INSTR' = 'knife'

Interestingly, Oswalt (1961: 267) reports that the plural of *ciwalaw* 'shirt,' which is derived from a complex verb meaning 'to crawl down' by means of the absolutive suffix (Oswalt's "literal" translation is 'crawling down'), was remembered by a consultant as having been *ciyalaw*, which incorporates additional verbal inflection for pluractionality, thus meaning something like 'crawling down repeatedly' when interpreted verbally. However, at the time of Oswalt's field work this form had been displaced by inflection with the regular nominal plural *ciwalá·ya*. Oswalt's (1961: 267) conclusion is that "[t]he Absolutive form may be used so predominantly as a noun that there is a tendency to lose sight of its verbal origin."

4.6.5.3.2. *Oneida.* There are four categories of nouns in Oneida (Abbott 2000: 47-48): (i) monomorphemic nouns with no internal structure, which are small in number, (ii) nouns consisting of a nominal root which require a prefix and a suffix that do not add any discernible meaning to the root, but instead simply serve to identify the forms in question as nouns, (iii) deverbal nouns derived from verb roots by suffixation of a nominalizing suffix (and in turn followed by a noun suffix), and (iv) "syntactic nouns," which are "formed by constructing a verb, typically with an indefinite pronominal prefix and serial suffix, and then simply using that verb syntactically as a noun" (Abbott 2000: 48). According to Michelson (1990: 76), "the majority of functional nominals are verbal derivatives or forms which are structurally indistinguishable from verb forms."

(65.) a. *kahnekóni?* / ka-hnek-No-ʔ/
'NEUT.AGENT-liquid/liqour-be.in.water/cook.in.water- STAT'
= 'wells, puddles'
b. *tetwaʔsʌ́thaʔ* /te-w-aʔsʌt-haʔ/
'DUALIC-NEUT.AGENT-drop:CISLOCATIVE-HAB'
= 'waterfalls, Niagara falls'
c. *tekaˑtʌ́heʔ* /te-ka-tʌ-heʔ/
'DUALIC-NEUT.AGENT-fly-HAB' = 'airplane'

In principle, the fourth way of forming nouns results in ambiguity as to their interpretation, that is, the forms may be interpreted either as verbs expressing actions or as nouns denoting entities (Michelson 1990: 77). However, Michelson (1990: 77-80) also gives a number of examples for different structural types in which the nominal reading is associated with a particular set of inflections, and when these are varied, the resulting interpretation is verbal and no longer nominal. More generally, she notes throughout that there is variation in how freely speakers provide "literal" meanings for the individual forms in question, which is further evidence that the forms are lexicalized with the nominal reading to different degrees.

In the domain of artifacts, it is particularly common to find the habitual aspect suffix combined with an instrumental suffix. As seen in examples (66c.) and (66d.), this combination of morphemes is also found in a number of body-part terms:

(66.) a. *yelathʌstákhwaʔ* /ye-lathʌst-hkw-waʔ/
'FEM.INDEF.SG.AGENT-get.something.up-INSTR-HAB' = 'ladder'
b. *yehyatúkhwaʔ* /ye-hyatu-hkw-aʔ/
'FEM.INDEF.SG.AGENT-write-INSTR-HAB' = 'pencil, pen'
c. *yewelaláhkwaʔ* /ye-wel-l-hkw-aʔ/
'FEM.INDEF.SG.AGENT-air/wind-be.in.or.on-INSTR-HAB' = 'lungs'
d. *-khwalákhwaʔ* /-khw-l-hkw-haʔ/
'-food-be.in-INSTR-HAB' = 'my stomach'

According to Abbott (2000: 53), one characteristic of Oneida terms referring to land forms is that they are either incorporated into a verb indicating position or stance, or they carry a locative suffix.[38] An example of the former construction from the database seems to be (67.):

(67.) *kélhiteʔ* /ke-lh-Nit-eʔ/
'ANOMALOUS.PREFIX-woods-be.in[39]-STAT' = 'forest, trees'

4.6.5.3.3. *Pawnee.* According to Parks (1974: 90), compounding is the most widely employed word-formation technique in Pawnee (see Parks 1974: 123-139 for a typology), although nominalization, which is at the focus of attention in this discussion, is also important.[40]

Nominal inflectional morphology in Pawnee is very sparse: there are locative, instrumental and diminutive suffixes; inflection for number is restricted to certain contexts, and only kinship terms are marked for (inalienable) possession. When not followed by any of the suffixes, and when not part of a larger construction such as a compound, nouns appear with the so-called "absolutive suffix" *-u* (Parks 1974: 97-98; this suffix appears as *-uʔ* in the lexical source). Other nominal stems cannot occur independently except when occurring with the diminutive *-kis*, of which Parks (1974: 94) says that it often seems to function as a plain nominal suffix without diminutive semantics. An example from the database would be *rikucki*, underlyingly /rikuc-kis/ 'bird-DIM' = 'bird' (compare Parks 1974: 103).

Nominalization in Pawnee occurs in a variety of subtypes, but all of them employ the subordinate form of the aspectual suffixes (subordination and aspect are intertwined in Pawnee, see Parks 1974: 188 for details). An example provided by Parks for a nominalization from an active verb (denoting actions and processes) involving the subordinate

[38] Parallels are also found in Tuscarora. For instance, the root *-(ę)ʔty-* 'bay' only occurs incorporated in the verb *-yę(T)-* 'to lay.'
[39] This root is featured in the term only "possibly" according to the consulted source.
[40] These statements are mirrored in the typological position of Pawnee: the majority of complex nouns in Pawnee are classified as being of the lexical type, and a minority as being of the derived type.

imperfective marker *-hus* is *kickawiriwu*, Parks's (1974: 117) literal translation of which is "paddles the water" (note the relevance of third person agreement being realized by zero). This is commonly used in terms for items of acculturation:

(68.) a. *ukaʔaatawiiriku ~ ukaatawiiriku* /ukaʔaata-wi-iirik-hus/
'be.a.shadow-LOC-see-IPFV = 'mirror, glass, window'
b. *rakcaakarikuku* /rak-caakarikuk-hus/ 'tree/wood-unlock-IPFV' = 'key'
c. *raktariihkaruukus* /rak-rariihkaraʔuk-hus/ 'tree/wood-write-IPFV' = 'pen'

Note that there is some variation as to whether the surface forms retain the final *s* of the aspectual suffix (Parks 1974: 117). The same is true of another type of Pawnee nominalization, the "gerundial" type (Parks 1974: 120; Parks and Pratt 2008 gloss the relevant prefix, *ra-*, as 'iterative'). Of this type, only one example is found in the database, *ra-raar-awarii* 'ITER-ITER-fly.around' = 'airplane;' the literal translation provided for this is 'one flying around.' Other gerundial nominalizations mentioned by Parks (1974: 120) also feature an aspectual suffix (see also Parks 1974: 121-122 for further structural variants of the gerundial type).

A further noteworthy feature of Pawnee is that a considerable number of terms for topological features and natural kinds lack a nominal equivalent entirely. Instead, the meanings are encoded formally as verbs. Examples include, but are not limited to:

(69.) a. *huu-kiihaar* 'enclosure-on.a.surface'
= 'be a valley, be a bottoms, as of a stream course'
b. *kic-takaahak* 'water-pass.down'
= 'water to drip, be a waterfall, be a rapid, rapids, as in a stream'
c. *tat-kus* 'stalk-to.be.sitting' = 'be a plant; plant to be (growing)'
d. *waa-wikiiʔac* 'hill-to.be.growing' = 'be a mountain, be a tall or high hill'

Another type that also occurs in other semantic domains consists of a verbal root with the nominal suffix *-uʔ:*

(70.) a. *kiwahaar-uʔ* 'be.a.lake-NOM'
= 'pond, lake, slough, body of water, area of standing water'
b. *awiriituʔ* /awirit-uʔ/ 'be.hot-NOM' = 'heat, steam'
c. *ratkahaar-uʔ* 'be.night-NOM' = 'night'

These are all pieces of evidence for a verb-centered nature of the Pawnee lexicon (see § 5.4.1. for further discussion with reference to Pawnee).

4.6.5.3.4. *Nuuchahnulth.* Relying on a large set of suffixes for the purpose of word-formation, this Wakashan language lacks compounding entirely according to Davidson (2002). Examples from the database are in (71.), see also Davidson (2002: 193-195) for an overview of devices and further examples.

(71.) a. *hita-ċus* 'LOC-dig' = 'bay'
b. *ƛ̓ap-ẏak* 'straddle-INSTR' = 'scissors'
c. *hapuuɬ* /hap-quw̌ɬʷ/ 'hair-on.the.face' = 'beard'
d. *ċaċałaqṅukum* /ča-ċałaq-nukʷ-im/
'RED-branch.out-in.or.at.the.hand' = 'finger'

Such formations can become more complex, as the examples in (72.) show.

(72.) a. *čaapaciiɬmaʕuk*
čapac-i:ɬ-maʕuk
canoe-making-one.skilled.in
'one who is skilled in making a canoe' (Nakayama 2001: 65)

b. *ċuuƛ̓ċuuyaẏak*
REDUP-ču-(y)a-ẏakʷ
ITER-wash-REP-instrument.for[41]
'basin' (Nakayama 2001: 66)

Nakayama (2001: 68) provides a detailed discussion of the scalar nature of conventionalization and lexicalization, which is similar to that noted for Greenlandic by Fortescue (1980), and of the varying degree to which Nuuchahnulth speakers are aware of the morphological structure of complex words, depending to some extent on the suffix they contain.

However, Nuuchahnulth also features clausal nominals with finite verb morphology. These are called "nouns derived from lexicalized verbs" by Davidson (2002: 328). Among the examples given by him are those in (73.).

(73.) a. *ḵʷiti·ḵʷitš* 'sticks on at intervals' = 'hummingbird'
(*ḵʷiT-* 'stick on' + iterative I aspect)
b. *qʷa·ya·ċikši·ɬ* 'Turns-into-Wolf' = '(man's) name' (*qʷayaċi:k* 'wolf' "with perfective (and hence verbalizing) suffix *-šiƛ̓* and iterative II aspect")

As Davidson (2002: 328) also notes, this type of nominal is a "particularly common source for personal and place names," and this fact explains their absence in the data for the present study. Notably, Nuuchahnulth is (in-)famous for an alleged lack of a distinction between noun and verb (a claim going back to Sapir 1911), and indeed Nakayama (2001: 44) emphasizes that Nuuchahnulth stems are generally able to predicate, regardless of whether they have a more "nouny" or "verby" semantics (compare the further discussion on distinguishing word classes in Nakayama 2001: 41-57). Nakayama (2001: 57) and Davidson (2002: 325) seem to converge in the position that the distinction between noun and verb

[41] Glosses : ITER 'iterative,' REP 'repetitive.'

may be made, but that it is grammaticalized only to a low degree, and of course, this fact is potentially highly relevant for the presence of nominals with finite verbal morphology.

The textual example in (74.) shows a derived noun for 'animal' embedded in natural discourse, and the derived noun in (75.) does the same for an expression meaning 'village' that is verbal in nature, as evidenced by the presence of inflectional morphology for mood (emphasis added in both cases).

(74.) *qʷiinuuƛ̓ḥiič* *wiiy̓aat*
qʷi-a·nu:ƛ̓-(q)ḥ-(y)i:-č wi:y̓a-'at
that.which-because.of-SIM-INDF.3-INF.3 never-SHIFT
why.you.would never

haẇaaḥin?at *saštup*
haẇ-a·ḥin-'at sa-štu·p
eat-deprive.of-SHIFT **crawl-species**
take.food.away.from animal
'The reason why you never take food away from **animals**.'[42]
(Nakayama 2003: 262)

(75.) *naʔaa ʔan ciiqciiqa ḥaa quuʔasm̓inḥʔi čaakupiiḥ*
naʔa· ʔan ciq-(y)a ḥa: qu:ʔas-m̓inḥ-ʔi· ča:kupi:ḥ
hear that speak-REP there human-PL-DEF men
hear that speaking there the.people men

ʔanč wim̓aaqƛ̓at ƛ̓awiičiʔat hiy̓atʔitq
ʔan-č wim̓a:qƛ̓-'at ƛ̓awa-'i·čiƛ̓-'at hiɬ-'atḥ-ʔi·tq
that-INF.3 incapable.of-SHIFT near-INC-SHIFT there-residing REL.3
that cannot.be.done get.close **where.they.lived**

qʷayaċiikštaqumɬ
qʷayaċi:k-štaqumɬ
wolf-groups
wolf.tribe
'He could hear the people saying that it was impossible to get close to the wolf **village**'[43] (Nakayama 2003: 157)

4.6.5.3.5. *Blackfoot.* In Blackfoot (Algonquoian), a similar but not quite identical situation is encountered. Here, it is possible to use intransitive verb stems with the appropriate verbal inflection as nouns by "reclassification" (Frantz 1991: 116), as in the following examples:

[42] Glosses: INDF 'indefinite mood,' INF 'inferential mood,' SHIFT 'perspective shifting,' SIM 'simultaneous ('while doing ...).'
[43] Additional glosses: INC 'inceptive aspect," REL 'relative mood,' REP 'repetitive aspect.'

(76.) a. *káta'yáípasskaawa* /káta'-á-ipasskaa-wa/
= 'NEG-DUR-dance-3S' = 'non-dancer'
b. *áókstakiwa* /á-okstaki-wa/
= 'DUR-read-3S' = 'reader' (Frantz 1991: 117, small caps added)

One difference to, for instance, Pawnee is that the third person affix is not zero, so that the clausal structure of these nominals is fully brought to light. As Frantz (1991: 116) points out, there is evidence for reclassification when these forms are pluralized: they take the plural suffix *-iksi*, which is used for animate gender nouns, but in the singular, the forms are indeed ambiguous since *-wa* occurs on both nouns and verbs.

Blackfoot, however, also features other types of nominalization. Deverbal nominals involve affixation of verbal markers indicating person and number that are "nearly identical" to those of an inflectional paradigm (the so-called conjunctive paradigm) found on verbs in subordinate clauses (Frantz 1991: 120). This is remarkably similar to the nominalizations found in Pawnee. As Frantz further points out, "[i]t is probably more realistic to speak of this phenomenon as clause nominalization, for not only do the verbs agree with subject or object, but all other elements which normally accompany verbs in clauses may be present with Conjunctive Nominals" (emphasis removed). However, it is important to emphasize that, unlike in other North American languages already discussed, the nominalized nature of such clauses is indicated by the presence of a nominal suffix following the verbal inflection from the subjunctive paradigm. The main types of nominals formed in this fashion are locational nominals, temporal nominals, instrumental nominals, and manner nominals, each indicated by different prefixes (Frantz 1991: 121-124). The most important constructions for the present study are locational nominals, formed by the prefixation of *it- ~ iit-*, and instrumental nominals, formed by the prefixation of *omoht- ~ iiht- ~ oht-*. Examples (from Frantz 1991: 121-122) include:

(77.) a. *iitáísóooyo'pi* /iit-á-iso-ooyi-o'p-yi/
'there-DUR-on-eat-21:NOM-IN.S' ='table (what one eats upon)'[44]
b. *iitáóoyo'pi* /iit-á-ooyi-o'p-yi/
'there-DUR-eat-21:NOM-IN.S' = 'where one eats/restaurant'
c. *kitsítáóoyihpoaawayi* /kit-it-á-ooyi-hp-ooawa-yi/
'2-there-DUR-eat(AI)-NOM-2P-IN.S'
= 'where you$_{2p}$ eat/your restaurant'
d. *iihtáóhpommao'pa* /iiht-á-ohpommaa-o'p-wa/
'INSTR-DUR-buy-21:NOM-3S' = 'money/what one buys with'

Comparing examples (77b.) and (77c.) indicates that they are apparently conventionalized as their meaning may be modified by the affixation of additional morphemes.

4.6.5.3.6. *Biloxi.* The "literal translations" offered in the consulted source for Biloxi suggest that clausal nominals with no nominalizing morphology are also present in the Siouan

[44] Glosses: DUR 'durative,' IN.S 'inanimate singular,' NOM 'nominal suffix.'

language Biloxi, in particular in neologisms in the domain of artifacts, as in the examples presented in (78.), which are presented together with the "literal" translation from the source.

(78.) a. *ina ´do^{n}-hi ´* 'clock' ("sees the sun")
b. *akŭtxyi´ on´ni ~ akŭtxyi´ on* 'pen, pencil' ("makes writing" or "makes books")
c. *o^{n} ´do^{n}ho^{n}ni ´* 'mirror' ("what is used for looking at or seeing")

The precise nature and extend of the phenomenon in Biloxi, however, cannot be determined on the basis of the extant sources.

4.6.5.4. Discussion: An areal or typological phenomenon?

The overall areal distribution of such formations remains at present somewhat unclear, although it is obvious that they are well entrenched in North America. However, Hagège (1993: 174) points out that "[m]ost languages can use whole sentences as if they were a single unit which functions like a noun" and cites English *forget-me-not* and French *m'as-tu-vu* 'lit. did you see me' = 'one who likes to show off' (next to the Mbum term for 'pinky' already mentioned in Hagege 1970) as examples. Given this evidence, the question one should probably ask about clause-like nominals appears to be one of quantity, that is, in which languages they form a regular mechanism of lexical enrichment and what enables them to do so rather than to ask for the necessary condition for such formations per se.

While the phenomenon is clearly well represented in North America, Beck (2005: 3) reports the following data showing a lexicalized noun with the structure of a relative clause from Upper Nexaca Totonac in Mesoamerica:

(79.) *ti:* *ki-ma:-w-í:*
HREL 1OBJ-CS-eat-CS[45]
'my wife' (lit. '3SG who feeds me')

It may be that kin terms, which are inherently relational and in this feature are unusual as nouns, are particularly amenable concepts to be expressed by such constructions. Thus it is unclear how far the phenomenon extends southwards in the Americas. Likewise, nominals with the surface structure of finite clauses are by no means restricted to North America. There are some particularly interesting cases reported from polysynthetic languages of Australia. In Bininj Gun-Wok, there is a highly polysynthetic verb complex with twelve prefix and two suffix slots (Evans 2004: 71), while the nominal morphology is restricted to noun class prefixes on some nouns and optional case suffixes for non-core grammatical roles (Evans 2004: 76). There is productive compounding, but with semantic restrictions as to which nouns are permissible elements of compounds. For instance, *bininj* 'man' or *daluk* 'woman' cannot be involved in compounding, and as a result compounds making reference to the occupation or typical activity of individuals are rendered impossible. Instead, "[c]oncepts of this type are frequently expressed as deverbal nominals, in which a fully

[45] Glosses: CS 'causative,' HREL 'human relative pronoun,' OBJ 'object.'

inflected word is simply used as a syntactic noun without any overt signaling of class-change; these may or may not have an incorporated nominal" (Evans 2004: 98). Examples can be found in (80.), also from Evans (2004: 98):

(80.) a. *ka-warlbu-n* '3-hunt-NP' = 'he hunts'/'hunter'
b. *kabarri-bolk-nahna-n* '3a-land-look.after-NP'[46]
'they look after the land' = 'land owners, custodians'

For Yir Yoront, Alpher (1991: 72) likewise reports the existence of "[n]ominalized clauses with no nominalizing morphology" that involve finite verbs. These are not very frequent, however, and are restricted to terms referring to animals and plants. Mithun (1984: 507) quotes the the example in (81.) from Berndt and Berndt (1951: 34-35) from Guwinggu, another Australian language which shows the same or at least a very similar structure:

(81.) *gandijigar'garmere gadbere ŋa:d nawu garigugbu´lere*
they keep it for us ours us these our bodies are dark
'they look after our (language) for us Aborigines.'

An obvious question that arises is: what is it that enables many languages of the Americas (and others, apparently particularly those of Australia) to coin words with clausal structures or at least resembling clausal structures to a considerably greater extent than, for instance, better known European languages? And, equally important, what allows these languages to not only coin such words, but also to have them enter the lexicon as conventionalized terms for a given referent? One factor that at least appears to play a certain role is head-marking clausal morphology for the conventionalization of the relative-clause-type of morphologically complex items exemplified by the Kiliwa terms constructed using *k*^{w}-. Although Kiliwa does have some elements of dependent-marking, including number marking (Mixco 2000: 15) and optional though common case marking indicating the grammatical function of NPs in the clause (Mixco 1965: 110), "[t]he noun phrase in Kiliwa is considerably simpler in its morphology than the verb phrase" (Mixco 2000: 24). A further relevant factor appears to be that the Kiliwa relativizer *k*^{w}- is merely an affix (on the verb, i.e., a head-marking element), which perhaps facilitates the conventionalization of complex word forms by virtue of the fact that, as an affix, it is a simple fixed unit within a lexical item, thus avoiding the structural clutter involved in other strategies. For instance, it is hard to imagine that a language with relative pronouns such as those found in languages of Europe would coin a complex description analogous to Kiliwa *(?)wa?kws?hin* 'house which runs' = 'automobile' that would then enter the lexicon as a fixed expression; rather, if at all, such a language likely would prefer other strategies of word-formation such as compounding to express the same content.

As for complex lexemes of the 'dances with wolves'-type, the discussion here must remain even more speculative, but an obvious area of grammar that is worth looking at is syntactic organization. One line of argumentation that is already found in Boas (1911)

[46] There is no index of abbreviations and glosses in Evans (2004).

claims that in some languages of North America overt noun phrases do not have argument status or that this status is at least questionable (see Milewski 1950 for an early comparative treatment of phrasal structure in languages of North America). Instead it is argued that they are in apposition to the pronominal affixes found on verbs, which constitute the "real" arguments of the verb (Jelinek 1984, cf. also Kibrik 1992: 137). This entails the verb opening up slots for arguments that are however immediately saturated by the pronominal affixes attached to them. This so-called pronominal argument hypothesis, however, is highly controversial. Connected to this question, but equally controversial, is the claim made for some languages of North America that they lack a lexically anchored distinction between nouns and verbs. For instance, this is argued for by Sasse (1993) concerning the Iroquoian language Cayuga; one piece of evidence provided is the existence of clausal nominals as in the discussion on Oneida above. Mithun (2000), in contrast, rejects this position, arguing that nouns and verbs can be distinguished by clear criteria in these languages, while still admitting that forms classified on morphological criteria as verbs may function syntactically as nouns. In her view, this fact does not entail the lack of a possible distinction between noun and verb (in both authors' discussions, the existence of nominals with verbal inflection is not the only piece of evidence adduced). For the purpose of the present study, it is neither necessary to accept or reject the pronominal argument hypothesis, nor to take a position on the question of the distinction between noun and verb in Iroquoian and other languages. With regard to the former, it is possible to at least note that noun phrases in Oneida and Pawnee, two of the languages in which the 'dances with wolves'-type of nominal is frequent, do not bear any overt signaling of their function in the clause, i.e. case (this is also true of Nuuchahnulth). This, to be sure, does not entail noun phrases not being arguments, but at least it is not incompatible with an interpretation of noun phrases as appositions rather than arguments. Still, it must be emphasized that, as seen above, there are structural differences in the formation of clausal nominals, and it would be necessary to assess the syntactic organization for each individual language in more detail before well-founded broad generalizations are possible. Looking at the relationship between referential and predicative expressions (as done by Sasse 1991) and the structure of discourse is perhaps also revealing, and will be done here using textual data from Oneida as an example.

In this study, individual lexical items are primarily considered in isolation, separated from the context they occur in. This is to some extent unavoidable, but one must not forget that their use by actual speakers is embedded into larger structures of discourse, which is in turn related to the overall lexicogrammatic system of the individual languages. Examining how the types of complex nominals are integrated into discourse is particularly instructive for languages such as Oneida, in which types of morphologically complex items exist that differ quite radically from garden-variety word-formation devices such as compounding and derivation. Consider the following excerpts from an Oneida story:

(82.) a. *nʌ ne? nʌ?n*
P P P

s-a-yakwa-an-itskw-A-hl-A-?tsl-ot-u-nyu-hE?
ITER-AOR-1EXPS-SRF-**thigh-JOIN-set-JOIN-NOM**-stand-DIST-DIST-SER
again we set up the **chairs**

tsi?	*yo-atek-ha?*	*wa?-t-yakwa-hwanhak-E?*
P	NO-burn-SER	AOR-DU-1EXPS-tie.up-PUNC
where	it (the fire) is burning	we tied around

'Then we would put the **chairs** all around the fire'[47]

(Michelson 1981: 13, 38-9, boldface added)

From the discourse-pragmatic point of view, reference is being made to 'chairs' on which something is predicated in this passage. The term for 'chair' in Oneida is morphologically complex, but conventionalized as a fixed expression, occurring in two alternants. When free-standing, its form is *anitskwahlákhwa?*. This is a morphologically complex artifact term of the type discussed in §4.6.5.3.2., featuring the instrument suffix *-hkw* followed by the habitual suffix *-ha?*. However, in Oneida, this type of "verby" nominal has the ability to be incorporated into a verb, in which case the sequence mentioned above is replaced by the nominalizing morpheme *-?tsl* (Michelson 1990: 76; this seems to be in contrast to other Iroquoian languages such as Cayuga, in which only monomorphemic nominal roots have the ability to act as incorporate, cf. Sasse 1993). It is this second alternant of the word for 'chair' that appears in the above passage incorporated into the verb root *-ot-* 'to stand,' which is in turn inflected multiple times with morphemes conveying mostly aspectual information, effectively creating a highly complex verb form consisting of a total of thirteen morphemes. What can also be seen in the above example is that the discourse in Oneida is heavily centered around the morphosyntactically defined verb (and presumably, Oneida would score quite low on referential density indices such as that developed by Bickel 2003). Although the discourse functions reference and predication can clearly be identified on the level of the sentence, morphosyntactically speaking, the stretch of discourse consists of a series of predicates realized by verb forms which are interspersed with unanalyzable discourse particles such as *nʌ, ne?, nʌ?n,* and *tsi?.* In other words, discourse is heavily verb-centered. In the following excerpt from the same story which further illustrates the nature of Oneida discourse structures, a clausal nominal with verbal inflection meaning 'quilt' occurs in an unincorporated form.[48]

[47] Glosses: 1EXPS '1exclusive plural subjective,' AOR 'aorist mode,' DIST 'distributive,' DU 'dualic,' JOIN 'stem-joiner,' NO 'neuter objective,' NOM 'nominalizer,' P 'syntactic particle,' PUNC 'punctual aspect,' SER 'serial aspect,' SRF 'semi-reflexive.'

[48] To be sure, these examples were selected to illustrate the types of constructions in Oneida that are of interest for the present study. As noted in the discussion, there are also other types of nouns in Oneida, such as those consisting of a root plus noun prefix and suffix, and of course they also figure in actual discourse. For instance, in the following excerpt, the noun root *-aht-* 'shoe' has a prefix indicating a female possessor and the noun suffix *-a?.*

b. *s-w-at-yel-ʌ* *kʌs* *yo-thole-ʔ* *thikʌ nʌs kwi*
ITER-NS-SRF-do-PERF P NO-be cold-PERF P P P
sometimes it is cold weather

te-ka-ihn-A-kkhani-ʔ *t-a-yukhi-yu-ʔ*
DU-NS-cloth-**JOIN-piece.together-PERF** CIS-AOR-F/ldp-give-PUNC
cloth pieced together she gave it to us

neʔ *waʔ-t-yakwa-at-haʔuwEʔek-Eʔ*
P AOR-DU-1EXPS-SRF-wrap.up-PUNC
this we wrapped it around us

'Sometimes it would be kind of cold and she would give us each a **quilt** to wrap around in'[49]

(Michelson 1981: 13, 39, boldface added)

One can again observe the same structures: a series of verbal forms (including that for the referential expression 'quilt') interspersed with a number of discourse particles. Similar remarks are also made by Mithun (1984: 505) for Mohawk, another Iroquoian language: "Normal discourse consists predominantly of morphological verbs, since verbs can function not only as clauses and predicates, but also as nominals, with no modification in form. Such nominals are often verbal descriptions of their referents, like *ra'swà:tha'* 'he extinguishes'/ 'fireman'" (see also further discussion in Mithun 1984 as to the role of lexicalization and to what extent such forms can be regarded as subordinate). The following example, which seems quite similar to the structures found in Oneida, gives a taste of this type of discourse organization in Mohawk:

(83.) *Ó:nen ki' ne rahtahkón:nis tahoná:khwe'.*
now just the he shoes makes he got mad
'At this point the shoemaker became angry'

(Mithun 1984: 506, quoting Phillips p.c.)

What is the relevance of this for the phenomenon at hand? The details of syntax and their interaction with the lexicon in Oneida (and Mohawk, and also other languages in which similar structures are found) clearly require detailed study by experts to elucidate the intricate relationship between grammar and lexicon. Nonetheless, it seems safe to at least

(i) *yah teʔ-w-eʔni-Ø* *n-aʔ-te-yo-stalathe-ʔ* *yako-aht-aʔ*
NEG NEG-NS-evident-PERF PART-AOR-DU-NO-shiny-PUNC FPOSS-shoe-NSF
'Her shoes were really shiny' (Michelson 1981: 25, 49)

Glosses: PART 'partitive,' FPOSS 'feminine-indefinite,' NSF 'noun suffix'

[49]Additional Glosses: CIS 'cislocative,' DP 'non-singular,' F/ldp 'feminine non-singular' (the sequence /l is unclear), NS 'neuter subjective.'

claim that the verb-centered nature of discourse allows for the smooth integration of clausal nominals, either in the form of relative clauses or, as is the case here, in the form of finite verb forms, into the overall organization of discourse. This statement, to be sure, is not an explanation of the phenomenon per se, but at least it helps to understand how the structures in question fit into the overall lexicogrammatical system of the language.

4.6.6. MIXED LANGUAGES OF EURASIA

4.6.6.1. Overview

Languages of Eurasia, including the Caucasus, have, as will be seen later, a comparatively high degree of unanalyzable, monomorphemic lexemes (attributed by Sasse 2001: 503 to the atypical millennia-long history of borrowing). With respect to the nature of their complex lexemes, it is also noteworthy that they are consistently inconsistent, that is, they mostly rely on a mixture of the derived and lexical strategy.

4.6.6.2. Greek

Modern Greek is typical in this respect. In this language, derived-type analyzable terms are found next to lexical-type analyzable terms, and it has been chosen to illustrate the typical mixture of the two types in Eurasia because many loanwords found throughout Eurasia are ultimately of Greek origin (as shown by example 84a. and b.), and while they are unanalyzable in the borrowing languages, they are analyzable in Greek. Examples of the different types (see Joseph and Philippaki-Warburton 1987: 216-229 for an overview of word-formation in Greek) are:

(84.) a. *oríz-ōn* 'bound/delimit-NOM.SG.NEUT' = 'horizon'
b. *mîlon toú Adám* 'apple-NOM.SG.NEUT ART.GEN.SG.MASC Adam' = 'Adam's apple'
c. *lik-ó-fōs* 'wolf-STEM.FORMATIVE-light' = 'dawn'

As examples (84b. and 84c.) also shows, complex lexemes may be of the compounding or phrasal type in Greek.

4.6.6.3. Laz

Another Eurasian language with this profile is Laz, a Kartvelian language. Analyzable terms are relatively few in number; if they occur, the lexical type is somewhat more frequent, but the derived type is also well attested:

(85.) a. *kinçi-toma* 'bird-hair' = 'feather'
b. *3'ari-gza* 'water-way' = 'river'
c. *o-ťoç-aṣe* 'DERIV-throw/shoot-DERIV' = 'weapon'
d. *o-bere* 'LOC-child' = 'womb'

While Laz has been in contact with Greek, it should be emphasized that the mixture of the derived and lexical type is frequent in Eurasia generally, in particular Western Eurasia. A mixed profile is also found in Basque, Bezhta, Sora, Welsh, and Kolyma Yukaghir. Likewise, Khalkha lexical items may be of both types as well (though here derived terms are a little

more numerous than in the aforementioned languages), and in Kildin Saami, analyzable lexemes in the database are of the lexical type, though generally Saami languages also feature a fair amount of derivatives.

4.7. CHAPTER SUMMARY AND A FIRST TYPOLOGICAL CORRELATE

In the above, a coarse first step was taken to the typologization of the dominant types of analyzable lexemes found in the world's languages. The only factor involved so far is the relative degree of analyzable terms of the lexical and derived type, and two extremes of the continuum on which individual languages may be placed were defined. For illustration, data from languages with different typological profiles and their interaction with the grammatical organization were presented. While this embryonic typology cannot (and is not intended to) capture particulars of all languages in the sample and does not assign every language to a particular type, it clearly does reveal an association of preferred word-formation techniques with certain linguistic types. It was pointed out that the areal distribution of those languages is skewed, with languages in which the derived type is dominant clustering in certain areas of the world, most notably the American Northwest (affixal type of polysynthesis) and the Northwest Amazon region in South America (derivational use of noun classifiers). However, as a first step, already at this stage of the investigation a striking grammatical correlate of this typology can be identified. In the discussion above, it was noted that several authors pointed to a predilection of so-called isolating languages for the lexical type, and an isolating language was defined rather loosely with reference to Greenberg (1960) as one with a low ratio of bound to free morphemes. However, it can of course also be defined with reference to individual grammatical properties. For instance, in a prototypically isolating language, arguments will not be marked for the type of relation they bear to the verb (case), and the verb in turn will not cross-reference properties of the arguments by means of affixes, while Vajda (2004: 421endnote3) notes that "[l]anguages with polypersonal verbs (i.e. verbs that internally cross-reference one or more syntactic terms or semantic roles in addition to the subject or agent) are normally regarded as polysynthetic on this basis alone."[50] Typically, this cross-referencing involves indication of person, and thus data from Siewierska (2005) on the occurrence of verbal person marking provide one opportunity to empirically measure and test the manifest hypothesis of whether there is a correlation between inflectional possibilities and a preference for either the derived or lexical type (note that this is only one test case, and the degree of synthesis cannot be exclusively defined by this measure, to be sure; examining many others categories in a similar fashion would be possible and indeed worthwhile). Siewierska distinguishes five types of languages: (i) those which do not have person marking, (ii) those that mark only the A argument, (iii) those that mark only the P argument, (iv) those that mark either A or P, and finally (v) those that mark both A and P. Since types (iii) and (iv) are quite rare cross-linguistically, the typology was simplified by removing these categories; additional data from the consulted grammatical sources was gathered for

[50] Whether or not this view is justified is another question; this statement is merely meant to illustrate the relevance of person marking as one factor for the isolating-(poly)synthetic distinction.

the languages in the statistics sample to fill gaps in the data because Siewierska's sample and the present sample do not always overlap (data are in Appendix C).

To ensure that the areal clusters of the types identified in § 4.3. do not interfere with the testing of this hypothesis, advanced statistical techniques are called for. These are introduced here, and will be used throughout the following chapters in various places. Generally, the data were analyzed by using Linear Mixed Effect Models (see Winter 2011 for a practical introduction) using R code by Bates and Maechler (2009) and Baayen (2009). This method of data analysis is becoming more and more important in psycholinguistics and phonetics research, when it is important to exclude unwanted factors, such as variation between test subjects, which have the potential to bias the outcome. The variable or variables of interest which one expects to have a systematic effect (in other words, which one believes to be a predictor) on some other variable (here, the percentage of derived terms) is typically included in the model as a so-called fixed effect, while a variable potentially having a random or unpredictable effect which one wants to control for is included as a so-called random effect (hence the term Mixed Model).

In this case, the variable of interest as a predictor is elaborateness of verbal person marking, and hence it is included in the model as a fixed effect, while area, in ensuring that it does not distort the obtained picture, is included as a random effect (see Cysouw 2010 for a similar approach in linguistic typology using Generalized Linear Mixed Models). As a rule, the Dryer-6 breakdown will be used exclusively to control for area in statistical modeling since it is the best-established of the different macro-areal regions of the world presently in use. The other two breakdowns as established by Nichols will be used in addition when areal effects on the typological variables under survey are of interest themselves in elucidating patterns both on a larger and a smaller scale that otherwise remain obscure under the Dryer-6 breakdown.

Normality and homogeneity of the model were then checked by visual inspections of histograms of residuals and plots of residuals against fitted values (and in addition, for further similar analyses, a Shapiro-Wilk test for normality and/or a correlation test was carried out between fitted and residual values if visual inspection leaves doubts as to the normality of the data). In this particular case, the response variable, i.e. the degree of derived terms, was subject to square transformation to ensure that the data are roughly normally distributed (this is always indicated when done in further tests to follow). All these are standard procedures with this type of statistical analysis to ensure that the observed effect really is genuine. Modeling was begun by including both a random intercepts component (which would correspond to differences in the value of the response in different areas, here, the areal asymmetries in the percentage of derived terms detected in § 4.3) as well as a random slopes component (which would correspond to differences in the impact of the predictor variable on the response in different areas, here, for instance, a strong effect of verbal person marking on the percentage of derived terms in one as opposed to a weak or even non-existent one in another area). Subsequently, a likelihood ratio test was carried out to ascertain whether the random intercepts component is needed and was removed if the likelihood ratio test did not reveal a significant difference between the models including and excluding it, which was the case for this particular model (unless otherwise indicated in the discussion of subsequent chapters, it was the case that

the random slopes component was not needed). The presented *p*-values are estimated by Markov Chain Monte Carlo simulations with 10,000 replicates (in case the resulting value for further models is between .03 and .07, that is, closely around the significance threshold, a further simulation with 100,000 replicates will be carried out to obtain a more reliable estimate).

A word of caution is necessary at this point: one of the most prominent traditional goals of linguistic typology is the identification of language universals and the determination of what is and what is not a possible human language. In doing so, many researchers not only seek to determine properties that living languages have, but also want to make inferences about any language ever produced by human beings in the past and any language that ever will be produced by human beings in the future. As pointed out by Maslova (2000), this poses a problem for any statistical analysis of the data, since the languages spoken today, from which such features can be extrapolated, represent only a small subset of the larger set of all languages ever produced, and their distribution is dependent to some degree on historical contingencies. Since no measures to remedy this fact (see Bickel 2008 for proposals) are applied here, the claims about correlations in the present sample are "universal" only in the sense that they extrapolate on the present-day (or near past) linguistic diversity in the world, and no claims are made that they constitute design features of all possible human languages (following Cysouw 2010: 258fn5 in this).

Finally, the result of the analysis is as follows: independently of area, there is a very significant correlation (p = .0022) between verbal person marking and the relative percentage of derived terms in the languages under investigation. As the plot in figure 6 shows, the derived type is least well represented in languages with no person marking, more frequent in languages in which the A is cross-referenced on the verb, and by far most frequent in poly-personal languages which cross-reference person information of both A and P on the verb.

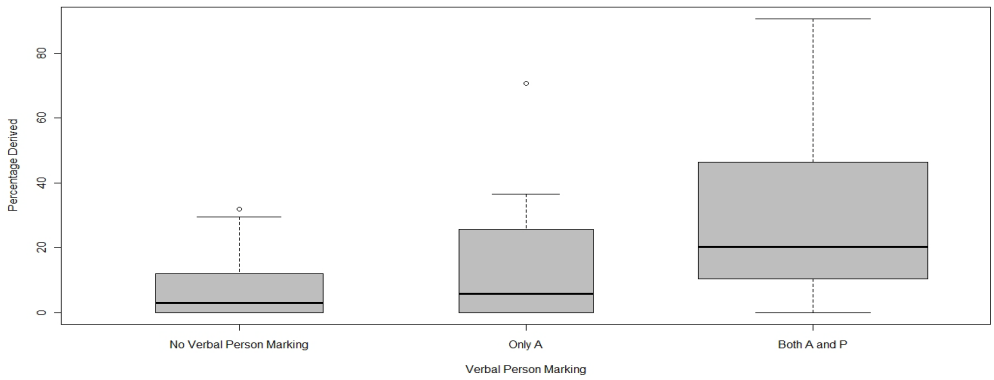

fig. 6: correlation between the preponderance of derived terms and verbal person marking

Consequently, even at this stage, a typological correlate to the classification can be added: THE MORE INFORMATION REGARDING PERSON IS CROSS-REFERENCED ON THE VERB, THE HIGHER THE AMOUNT OF TERMS OF THE DERIVED TYPE FOUND IN THE LEXICON. This fact can be interpreted as evidence in favor of the statistical correlation between an "isolating" typology and the dominance of compounding as the main word-formation strategy (as well as evidence in favor of a correlation between an elaborate verbal morphology with the derived type). [51]

Nonetheless, the above typology does not explain one of the major factors of interest for this comparative study of analyzable terms in the vocabulary, namely the differential degree to which the lexicon actually features such terms: from only a very few in some languages to quite impressively large numbers in others. There is no statistical effect at all of a language's preference for analyzable terms of the derived or lexical type on the actual number of these that are found in the lexicon ($p = .8610$ by the same Mixed Model design with percentage of derived terms as the response square-transformed). Thus, languages of both types may feature either a rather small or a comparatively high number of analyzable lexical items, regardless of their type. This means that a new dimension must be added to this typology, and other factors that influence this variable must be sought. In addition to considering other quantitative evaluations of the observed data, this is the question that will be the main focus of the following chapter.

[51] But note that there is no evidence for a correlation between the presence of case-marking for core arguments with the placement of languages on the derived-lexical continuum.