

CLIPP

Christiani Lehmanni inedita, publicanda, publicata

titulus	Roots, stems and word classes
huius textus situs retis mundialis	http://www.christianlehmann.eu/publ/lehmann_roots.pdf
dies manuscripti postremum modificati	01.11.2007
ocasio orationis habitae	Conference on Universality and Particularity in Parts of Speech Systems, Amsterdam, 8.-9.6.2006
volumen publicationem continens	<i>Studies in Language</i> (Special Issue)
annus publicationis	2008
paginae	546-567

Roots, stems and word classes

Christian Lehmann

University of Erfurt

The great tragedies of science are the slaying of beautiful hypotheses by ugly facts. (Thomas H. Huxley)

Abstract

The assignment of a linguistic sign to a word class is an operation that must be seen as part of the overall transformation of extralinguistic substance into linguistic form. In this, it is comparable to such processes as the transitivity of a verbal base, which further specifies a relatively rough categorization. Languages differ both in the extent to which they structure the material by purely grammatical criteria and in the level at which they do this. The root and the stem are the lowest levels at which a linguistic sign can be categorized in terms of language-specific structure. Further categorization is then achieved at the level of the syntagm.

An empirical investigation comparing the categorization of roots and stems in a sample of six languages (English, German, Latin, Spanish, Yucatec Maya, Mandarin Chinese) turns up far-reaching differences. These differences in the amount of categorization that languages apply to linguistic signs at the most basic levels throw into doubt any thesis claiming universal categoriality or acategoriality for roots. Such a static view must be replaced by a dynamic one which asks for the role of categorization in linguistic activity. At the same time, these differences raise the issue of the amount of structure – or of grammar – that is necessary for a human language.¹

1. Introduction

The Modist theory of the parts of speech as laid down by Thomas of Erfurt (14th cent.; cf. Bursill-Hall 1972) goes roughly as follows: Pre-linguistic concepts are like an amorphous substance that gets a linguistic form by being assigned a grammatical category. The most important of these categories are the parts of speech. They have a basis in perception, but do not directly reflect it. Instead, the transfer of a concept into such a category is an operation of the intellect. The notional aspect of the operation is the addition of a *modus significandi* (roughly, the meaning of a grammatical category) to the basic concept. Its formal aspect is the transfer of a *dictio* – roughly, a root – into a *pars orationis* ‘part of speech’.

¹ Thanks are due to three anonymous reviewers for helping me improve this paper.

Similarly, Coseriu (1955) considers that the meaning of a lexeme is composed of its lexical meaning (“semanteme”) and its categorial meaning (“categoreme”). The former concerns what is signified, the latter concerns how it is signified. Languages differ in the degree of “autonomy” of the lexical meaning as against the categorial meaning, which essentially means that a given lexical meaning may or may not entail a particular categorial meaning in the language. This autonomy of the lexical meaning is greater in English and Chinese than in Spanish (Coseriu 1955, § 2.3.2) and “most Indo-European languages” (§ 5.2.2). To give an example from the data of the present study to illustrate what Coseriu means: The concept ‘comfort’ is coded in Spanish by the stem *consol-*, which can only be inflected as a transitive verb. That is, given the lexical meaning as paired with a stem, the word class is given, too. The same concept is coded in Mandarin Chinese by the stem *ānwèi*, which can be used as a verb (‘to comfort’) or as an adjective (‘comforting’). Thus, the lexical meaning of the Mandarin stem is more independent from specific word classes than in Spanish. In Coseriu’s theory, the categories are not syntactic classes, but categories of speaking, in a sense to be made precise in the next section. In general, however, languages do not contain “pure” lexical meanings that would be associated with a category only in speaking. For every semanteme, one of its alternate categories is usually primary (p. 39f).

The above are postulates of linguistic theories. As such, they cannot be upheld; we will come back to this in § 4. In what follows, we will consider them as hypotheses. The central hypothesis in this connection is that roots are precategorial. It will be tested on samples of stems and roots taken from six languages. In § 2, some basic notions of grammatical categorization are introduced. The central section of the paper is § 3, which first discusses the methodology to be employed and then presents the results of the analysis. § 4 draws some theoretical and methodological conclusions from the findings.

2. Grammatical categorization

The question of grammatical categorization in the world’s languages has often been analyzed within the confines of word-class systems. It has been observed that languages differ in the number and kinds of word classes that they distinguish and that word-class distinctions seem to be stricter in some languages than in others. In this perspective, it would appear that languages with many different word classes and with sharp boundaries between them fulfill some function neglected by languages with fewer or more fluid word classes. Here we will widen the perspective a bit and view the categorization of lexemes into word classes as one kind of **grammatical categorization**, where ‘grammatical’ is taken to comprise ‘syntactic’ and ‘morphological’.

It is a theorem of semiotic theory that linguistic expressions must be composed of meaningful elements that belong to categories. This should be deducible from the requirement that there must be compositional rules of forming the meaning of complex expressions, which in turn follows from yet more basic goals of cognition and communication. Taking this for granted, we may ask at which level of grammatical structure such categorization takes place. I am here referring to the **complexity levels** of grammatical structure. The levels relevant for the categorization of units with lexical meaning are the ones of T1:

T1. *Levels of grammatical categorization*

phrase

word form
stem
root

Consider E1 for illustration:

E1. The undermining of established theories is forbidden.

- *Mine* is a lexical root that belongs to either of the categories noun or verb.
- *Undermine* is a stem that belongs to the category verb.
- *Undermining* is a word form that belongs to any of the categories noun, adjective or adverb (the latter when heading a gerundial).
- *The undermining of established theories* is a phrase that belongs to the category noun phrase.

E1 refers to what the noun phrase designates and predicates something about it. The propositional operation of reference works with an expression belonging to the category noun phrase. This operation does not care for the category assignments of the lower levels. In this perspective, the traditional name of the categories here at stake, viz. ‘parts of speech’, is entirely apt because it is at the level of the propositional operations that categorization is needed. Categorization at lower levels, e.g. the level called ‘word classes’ or ‘lexical categories’, is required only to the extent that compositional semantic complexity below the level of propositional operations is wanted. That is to say: Speakers certainly need semantic complexity below the level of propositional operations. However, as word-formation shows, it need not be compositional.

A meaningful expression may be categorized at some lower level and then be handed through to the upper levels with its category unchanged, as is the case for the subject of E2 (cp. with the subject of E1).

E2. War is forbidden.

Assigning an element to a target category or equipping it with such a category is an operation that can take place at various levels. The element may be stored in the inventory under the target category or under a different category or without any category. In the first situation, the operation of categorizing the element has, so to speak, already been done in the language system. The category need not be marked on the element if it is part of the element’s identity. For instance, the English stem *war* belongs to the category noun, and that is part of its lexical information; but the fact is not marked on it by any structural means. In the second and third situations, the target category will somehow be coded on the element or in its immediate context (cf. Kuryłowicz 1936). For instance, the assignment of *undermining of established theories* to the category noun phrase is marked by the formative *the* appearing in front of it. These options are available to languages; and they may differ typologically in resorting to one rather than another solution of the categorizing task under different conditions.

The categories we are talking about are **structural categories**. They are, thus, part of the particular language system. They are distinct from the **cognitive categories** into which signs fall. ‘Property’ is such a cognitive category, and both the adjective *clever* and the noun *cleverness* fall into that category. We will come back to the question of how the cognitive category becomes relevant for the grammatical categorization.

Before a sign reaches the level of the utterance, it may be categorized and recategorized several times. With reference to the levels of T1, we may distinguish between the primary and the final categorization of a sign. The **primary categorization** is the one at the lowest level, the **final categorization** is the one at the highest level. Examples of multiple recategorization are not hard to come by; one was already mentioned in E1. For subsequent discussion, it will nevertheless be sufficient to work with the primary and final categorizations as a binary distinction.

In a way to be made precise by linguistic theory, the passage from the lower to the higher levels of grammatical organization up to the text level is associated with the transition from *langue* to *parole*. The latter is a transition from the virtual to the actual and is, therefore, not gradual like the passage from the lowest to the highest level of T1. However, as is well known from research on grammaticalization, the language system determines structure more rigidly at the lower levels of T1 than at the higher levels. At the highest level, the speaker freely selects and combines linguistic units; the lower the level of complexity, the more he can rely on ready-made units of the language system, which predetermines their structure. Therefore, the primary and the final categorization of linguistic units are determined at different levels, by cognitive and communicative principles of a different nature.

What determines the final categorization of an expression?

The final categorization of a sign is determined by the syntactic function it has to fulfill in the sentence. That, in turn, is determined by the propositional operation (reference, predication, modification; cf. Croft 1991) to be performed on it. In terms of a teleonomic hierarchy, the speaker chooses his means according to his cognitive and communicative goals. Sometimes he has to adapt the means that the system offers him. That is to say, the use of a certain expression in a certain category may necessitate some prior operation of recategorization on it.

What determines the primary categorization of an expression?

While the syntactic function of an expression is the only factor in its final categorization, it cannot be the only factor in the primary categorization, because a lexical concept² must be available for different syntactic functions so that the primary categorization cannot foresee the ultimate use that a sign will be put to. Primary categorization has to proceed by a probability calculus of the sort: what will most probably be the syntactic function of this lexical concept? For this decision, there are by definition no other criteria available than the meaning of the sign. Primary categorization of lexical concepts is essentially determined by universal cognitive principles. The best-understood of these rely on the time-stability of a concept (cf. Croft 1991):

- If the concept has a high time-stability, i.e. it is thing-like, then it will be used in referring. The category of a word whose primary use is reference is ‘noun’.
- If the concept has a low time-stability, i.e. it is event-like, then it will be used for predication. The category of a word whose primary use is predication is ‘verb’.

If the concept is not clearly determined in terms of time-stability, then a couple of other major classes become available. Some of these, above all the adjective and similar categories, are still

² The expression ‘lexical concept’, though well-established in current linguistics, begs an important theoretical issue: Concepts as such are language-independent; but the assignment of a given concept to either lexicon or grammar is a matter of language-specific structure, more precisely, of the linguistic operations/processes of lexicalization and grammaticalization. There is nothing in the nature of a concept that would make it intrinsically lexical or grammatical.

weakly determined by time-stability, viz. by the medium stability characteristic of properties and states. Here, however, categorization becomes more arbitrary and more language-specific. Little is known, *a fortiori*, about the factors underlying the primary categorization of concepts as adverbs.

In this paper, only the categorization of lexical concepts in terms of parts of speech is dealt with. There is, on the other hand, the set of categories of grammatical formatives, variously called ‘secondary grammatical categories’ or ‘morphological categories’. The assignment of a concept to any such category is a matter of grammaticalization, not to be treated here.

3. Categoriality of roots and stems

What I am presenting here is an interim report on ongoing research. Its point of departure is the pre-theoretical observation that there appear to be substantial differences among languages with regard to category determinacy of roots, i.e. to the extent to which roots are categorized at all. Instead of category determinacy or categorial specificity, we will say ‘categoriality’ for short. The questions to be answered by empirical investigation are the following:

- What kind of variation is observed in the primary categorization of lexical concepts?
- If there are significant differences among languages, with what do they correlate?³

A small convenience sample of six languages was taken, including Latin, Spanish, English, German, Mandarin Chinese and Yucatec Maya. Some of these languages were included because there are long-standing suppositions that they behave specially with respect to the two leading questions;⁴ others were included because data were easy to come by. The inclusion of both Latin and Spanish renders the diachronic perspective possible, which will occasionally be taken here.

3.1. Methodology

The object of the present research is, thus, the set of roots of a language, and the task is to determine the range of word classes for these. Since the roots of a language number in the thousands, one may wish to take a sample. One method that would, in principle, suit the goal is to take a random sample from a dictionary. However, since in comparative research on grammar and lexicon, an onomasiological perspective must be taken (cf. Lehmann 2005), the choice method here is the selection of a lexical field. A field is wanted which is liable to present variation in categorization. Variation in two respects is needed: Different word classes should be represented in the field; and average categoriality of items in the field should be low. (These two kinds of variation are expected to correlate.) These conditions rule out lexical fields of physical objects or of destructive acts. In such fields, the cognitive factor determining primary categorization in the sense of § 2 is relatively strong. That means, most lexical concepts concerning physical objects in any language will be categorized as nouns; and most lexical concepts concerning destructive acts will be coded by transitive verbs. Here universal cognitive principles of categorization prevail, with the result that

³ This question has seldom been addressed in the field of word classes, a recent exception being Haig 2006, whose methodology relies on word-formation. The methodological approach taken here partly resembles the one taken in the research on causative and anticausative coding of a set of verbal concepts in various languages as executed in Haspelmath 1993.

⁴ For instance, Coseriu’s (1955) above-mentioned statement on Chinese is part of a century-old debate; cf., e.g., Simon 1937 reacting to a verdict in Misteli 1893 that Chinese did not have word classes.

root categoriality will be relatively high and typological differences among languages have little chance to assert themselves.

A field that would qualify well by these conditions would be sounds and noises. This, however, presents a practical problem of getting data. For lesser-known languages, dictionaries do not suffice in this domain. For living languages, data may be obtained in fieldwork, but the diachronic perspective will be largely excluded, and questions concerning changes in the primary categorization of lexical units could not be posed. I therefore chose the domain of experience, which includes the subdomains characterized in T2 (cf. Verhoeven 2007):

T2. *Conceptual domain of experience*

subdomain	definition	subdivision	example root concepts
bodily sensation	feelings related to the experiencer's body	saturation, temperature, irritation, health condition	hunger, tire, itch ...
emotion	cognitively based feelings	self-directed, other-directed	happy, angry, fear ...
volition	psychological energy determining actions	psycho-physical disposition, intention	want, greed, hope ...
cognition	internal experiences based on mental functions	presence or absence of information, conceptual activity, propositional attitudes	know, dream, decide ...
perception	input of information through the senses	sight, hearing, touch, taste, smell	see, loud, noise ...

I only considered the first four of the five subdomains of T2, the reason being again an expectation that in the remaining subdomain, perception, cognitive principles will determine categorization to a large extent.

A set of basic concepts in these subdomains was established on a rather intuitive basis (s. § 4 for the obvious methodological weakness involved), but exclusively by cognitive, not by structural criteria. In this approach, the English words *clever* and *cleverness* code the same concept. The full set of concepts is enumerated in T3.

T3. *Basic concepts*

subdomain	concepts
bodily sensation	feel, rest, hunger, thirst, satisfied, enjoy, refresh, cold, cool, hot, chill, sweat, wake, brisk, sleep, tire, feeble, flabby, sound, intact, recover, sick, ill, suffer, endure, pain, ache, itch, numb, cramp, tickle, wanton, lust, rut, irritate
emotion	temper, calm, happy, bliss, love, proud, comfort, startle, astonish, scare, dread, shudder, terror, timid, shy, fear, anxiety, excite, fury, rage, grudge, wrath, annoy, anger, vex, resent, angry, bother, bore, patient, disgust, sad, grief, sorrow, shame, serious, serene, gay, please, amuse, frenzy, enthuse, charm, hate, envy, pity, move
volition	want, wish, strive, endeavour, effort, instinct, long, yearn, desire, greed, eager, wait, hope, inclined, mood

cognition	mind, soul, think, know, acquaint, presage, understand, decide, puzzle, mad, remember, learn, forget, imagine, dream, respect, honor, believe, trust, agree
-----------	---

These concepts are translated into the target languages. (The data and subsequent analyses are not reproduced here for want of space; a sample is shown in the appendix.) No attempt is made to obtain precise translation equivalents, because that is both impossible and inessential to the task. Instead, to the extent that there is a choice among expressions in a language, preference is given to morphologically simple expressions whose literal meaning is in the lexical field, so that the root itself is an experiential root rather than functioning in an experiential expression by metaphor. The condition of simplicity makes us choose *satisfy* as the English coding of the concept ‘satisfied’ instead of the participle *satisfied*, and similarly *anger* as the English coding of the concept ‘angry’.⁵ The condition of literality makes us exclude English *upset* as a coding of ‘angry’. Sometimes this condition cannot be satisfied; for instance, *move* is the only straightforward English coding of the concept ‘move emotionally’ (German *rühren*). If two concepts are rendered by the same stem in the target language, the stem enters the sample only once, and accordingly the sample of stems will be smaller for that language. For instance, I could not find in Spanish a distinction between ‘timid’ and ‘shy’.

Given a concept and a root coding it, then there are often more than one stem that transfers the root into different word classes, as in English *anxi-ous* and *anxi-ety*. If the above criterion of simplicity does not decide it, one of the stems is chosen arbitrarily. If different concepts are represented by different stems based on the same root, as in Latin *sentī-* ‘feel’ and *ad-sentī-* ‘agree’, then these stems are included. Such discretionary decisions do no harm as long as we do not ask for the particular word classes that prevail among the stems in the field; this latter question would indeed require a refinement of the methodology.

Then each expression is analyzed morphologically, and the root is identified. Items based on non-native roots are excluded if there are alternatives. Finally, the set of lexical categories for each root is determined by criteria dealt with in § 3.2. Here again, the degree of methodological refinement is low. The most important word classes are the same for the six languages, viz. noun (N), verb (V) and adjective (A). The concepts of our field are practically never coded as adverbial or adpositional roots, although roots of these categories may appear in compound stems. Some languages have roots of minor categories. Again, these differences are not important since the epistemic interest is not in the nature of the categories, but in their distinction.

3.2. Analytic decisions

In what follows, some principles that are relevant to the morphological analysis, in particular in the identification and categorization of roots and the distinction between root and stem, are formulated.⁶

⁵ The direction of derivation is determined by purely structural, not by historical, psychological or usage criteria. Thus, for example, the basic term coding the concept ‘tired’ in English is the verb *tire*, no matter how much other criteria might speak in favor of *tired*.

⁶ No confusion should arise from the fact that the expressions ‘categorization of linguistic units’ or ‘determination of the category of a linguistic unit’ are ambiguous, because they may refer to a theoretical or to a methodological aspect of the object. The theoretical question is what factors in language activity determine the assignment of a linguistic unit to a particular category. This has been discussed briefly in § 2. The methodological question is how the linguist

The methodology involved in identifying the word classes that a given item belongs to will be taken for granted here, although it is currently the object of heated debate (see *Linguistic Typology* 9/3, 2005).

A root is assigned to word class W if it occurs in the contexts that define W. For inflecting systems, that means that the root combines with inflectional morphemes of W. For non-inflecting systems, the context is a syntactic construction. In both cases, it is, of course, required that the root be insertable in its context without further morphological or phonological modification, especially without any derivational morphemes or thematic suffixes. This condition assures, e.g., that German *Wut* ‘fury’ is only categorized as noun. The verb *wüten* ‘rage’ has a stem *wüt-*, which differs from the root by metaphony.

By this criterion, a root may belong to more than one word class simultaneously (e.g. Engl. *chill* is A, N or V), or it may belong to no word class (e.g. Engl. *aggress* as in *aggressive*). In the latter case, it is assigned to category X. These decisions presuppose semantic sameness. For instance, ‘mind’ is *mente* in Spanish. The root is *ment-*. This root cannot directly inflect, i.e. it belongs to category X. The fact that there is a (homonymous) root *ment-* which does directly inflect in *mentir* ‘lie’ plays no role in this assignment.

A **recurrent morpheme** is an association of a significans with a significatum that recurs in different contexts. By the principle ‘once a morpheme – always a morpheme’, a stem containing a recurrent morpheme is considered complex, no matter whether the rest is recognizable as a recurrent morpheme. This is relevant in two cases:

- Unproductive derivational operator, as in Yucatec Maya *uk’-ah* (drink-?) ‘thirsty’. Such cases present no problem of categorization, since the category of the root can be ascertained in other contexts.
- Unproductive root, as in Yucatec Maya *k’oh-a’n* (?-RSLTV) ‘ill’. In such cases, the root is classified as X (despite the fact that in productive formation, the resultative suffix only combines with a verbal base⁷).

A morpheme or paradigm that is used productively to convert stems into category W is or involves a derivational operator; i.e. it is not (merely) an inflectional morpheme or paradigm, even if it also appears on roots that only belong to W. This concerns the Latin **thematic vowels**. These function in transcategorization, e.g. *cal-e-re* ‘be hot’ as opposed to *cal-idu-s* ‘hot’; but they also appear on many verbs and nouns that have no counterpart in another word class, e.g. *par-e-re* ‘obey’. Thus, by this principle, the Latin thematic vowels are derivational operators; they form neither part of the root nor part of the inflectional ending.

In many languages, every **adjective** can be used as a noun. Given a root used as an adjective, then its use as a noun is considered an alternate root categorization if it is an abstract noun (e.g. German *stolz* ‘proud’, *Stolz* ‘pride’), but it is considered a recategorization at the syntactic level if the noun

finds out what category a certain linguistic unit belongs to. That involves methods of structural analysis like those mentioned here. The two questions are entirely independent of each other.

⁷ Assignment of *k’oh-* to category V would not be based on any independent evidence, but just on analogy. However, since the root is no longer in use, the formation is not a product of today’s language system, but of an earlier stage, where *-a’n* may well have combined with bases of different categories. Anyway, as explained at the end of § 3.1, it does not affect the present research question if such decisions have to be revised.

designates an object (typically, a person) that the meaning of the adjective applies to (e.g. Engl. *young*). In other words, *stolz* is categorically indeterminate (A/N), but *young* is not (A).⁸

For Spanish nouns, the suffixes *-o*, *-a* and *-e* are considered derivational suffixes, as their Latin predecessors. As a consequence, the root *deleit-* is a verb root (*deleit-ar* ‘delight’), but not a noun root (*deleit-e* ‘delight’ is derived). In the adjectives, instead, the endings *-o* and *-a* are inflectional endings (masculine vs. feminine); consequently, what remains after their subtraction is an adjective stem. Similarly, the thematic vowels of the three conjugation classes (*-a*, *-e*, *-i*) are considered part of the inflectional endings, not as derivational operators. This means that roots like *salv-* ‘save, safe’ can directly inflect (in two different categories, in this case). These latter analytic decisions differ from the corresponding decisions taken for Latin. This difference makes Spanish appear as a language of a slightly different morphological type than Latin. The decisions themselves are by no means cogent; but as we shall see in § 3.4.1, they seem to have little influence on the results.

3.3. Calculus

On the basis of the analysis described in the preceding sections, each item – root or stem – in the sample was assigned a numerical value for categoriality, by the following consideration: The lower the number of categories that an item may be used in, the higher its categoriality. Consequently, its categoriality is, in principle, the reciprocal value of the number of its categories (s. T4, rows 1 – 3). Only if the item can be used in no category, its categoriality is stipulated to be 0.⁹ The following values were used in the data analysis:

T4. *Values of categoriality*

number of categories	value
1	1
2	0.5
3	0.33
0	0

Moreover, the number of morphs composing a stem was noted. T5 is an illustrative section of the table of English stems:

T5. *Values of some English stems*

stem	categoriality		morphs
	categories	value	
re-cover	V	1	2

⁸ This is not to deny the difference between conversion of an adjectival into a noun phrase at the syntactic level, analyzable as its combination with a zero head, and conversion of the word-class by substantivization of an adjective at the level of word-formation. However, the criteria for such a distinction refer to the use of adjectives in texts, whereas here we are dealing with lexical inventories.

⁹ This decision is less than satisfactory, since by the logic of the reciprocal value, the categoriality value of an item that can be used in 0 categories should be ∞ . This figure, however, would have ruined all of the calculations. Nor would it make sense to say that an item that can be used in no category has a higher categoriality than an item nailed down on one category.

sick	A	1	1
ill	A	1	1
suffer	V	1	1
en-dure	V	1	2
pain	N/V	0.5	1
ache	N/V	0.5	1
itch	N/V	0.5	1

Here is the corresponding section for the roots appearing in the sample of T5:

T6. *Categoriality of some English root types*

root	categoriality	
	categories	value
cover	X	0
sick	A	1
ill	A	1
suffer	V	1
dure	X	0
pain	N/V	0.5
ache	N/V	0.5
itch	N/V	0.5

As said in § 3.2, homonymy is paid attention to, so that the morpheme *cover* that appears in *recover* is distinct from the morpheme *cover* ‘wrap’. The former is not usable as a stem and therefore receives category X and value 0.

As explained in § 3.1, the samples of stems of the six languages differ in size. Since in each language sample, each stem appears only once, the number of stem types equals the number of stem tokens. This is not so for the morphs – roots and bound morphs –, since the same morphs may be used in different stems. For instance, *re-* occurs in *resent*, *remember* and *respect*. Therefore, for each language an inventory of the (types of) major class roots appearing in the sample was made.

On the basis of these data, for each language sample figures F1 – F5 were computed as integers:

- F1. stems
- F2. morph tokens
- F3. root tokens
- F4. root types
- F5. non-root tokens: F2 - F3.

The morphological complexity of stems was determined by two figures:

- F6. morphs per stem: F2 / F1.
- F7. non-roots per stem: F5 / F1.

The average categoriality was computed for stems and for roots of each language as follows:

- F8. categoriality of stems: arithmetic mean of categoriality values of stems (a segment of which is illustrated in T5, col. 3), i.e. sum total of these categoriality values divided by F1;
- F9. categoriality of roots: arithmetic mean of categoriality values of root types (cf. T6, col. 3), i.e. the sum total of these categoriality values divided by F4.

For both of these means, the tables below also show the standard deviation. This, however, will not be analyzed since these distributions differ considerably from normal distributions.

3.4. Results

3.4.1 Categoriality of stems

T7 shows the categoriality of stems (F8) for the six samples:

T7. *Categoriality of stems*

stems language	types = tokens	mean categoriality	standard deviation
Latin	107	0.99	0.068
Spanish	116	0.98	0.102
Yucatec	78	0.96	0.144
German	116	0.86	0.230
Mandarin	112	0.84	0.233
English	116	0.77	0.245
average		0.90	

As appears from T7, there are considerable differences among the sample languages in this respect. At one pole of the scale, we have Latin and Spanish, where essentially every stem belongs to just one word class. In Latin, a stem that can be inflected for more than one word class is the exception. Our sample contains two of them, *met-u-* ‘fear’ and *ir-a-* ‘wrath’, which may be declined or conjugated. The fact that Spanish is so close to Latin is astonishing given the analytic decision taken in § 3.2 to the effect that thematic vowels of adjectives and verbs are considered as part of the stem in Latin, but not in Spanish. This may be interpreted to mean that high categoriality of stems is a typological feature of both languages, although they differ in their techniques to achieve it. At the opposite pole, there is English, where roughly every second stem belongs to more than one category. Expectably, Mandarin is close to the lower pole, too, but as has been remarked repeatedly by Sinologists, its stems are not as indeterminate as the English stems are.

For a language such as English, with low stem categoriality, the speaker’s task of assigning the words to categories is fully achieved only at the level of syntax. This is done by inserting words into certain syntactic templates which force a syntactic category on them. In a language with high stem categoriality, such as Latin, the syntax contributes nothing to the categorization of words, which means that such templates play a minor role in constructions.

3.4.2 Categoriality of roots

T8 shows the categoriality of roots (F9 of § 3.3) for the six samples:

T8. *Categoriality of roots*

language	roots	types	mean categoriality	standard deviation
German	117		0.78	0.327
Yucatec	71		0.76	0.404
English	116		0.62	0.381
Mandarin	155		0.60	0.465
Spanish	111		0.59	0.490
Latin	86		0.30	0.462
average			0.61	

This time, the differences among the languages are even more striking. At the upper pole of the continuum, there is German, where roughly every second root comes with a specification of its unique word class. Another language whose root categoriality is well above average is Yucatec Maya. This finding is in consonance with earlier work highlighting the rigid grammatical relationality of Yucatec roots (Lehmann et al. 2000; cf. also Lois & Vapnarsky 2003).¹⁰

At the lower pole, we have Latin, where the root that may directly be inflected according to a word class is the exception. Relevant examples are *prem-* ‘press’ and *ang-* ‘frighten’, which are verb stems, and *felic-* ‘happy’, which is an adjective stem.¹¹ Much more commonly, a root is first extended by a thematic vowel before it can inflect.

Although Spanish has the second-lowest value in this sample, Spanish roots are much more category-specific than Latin ones. This difference has nothing to do with the analytic decisions explicated in § 3.2, since in neither language are the thematic vowels considered to be part of the root. In a diachronic perspective, it may be interpreted to mean that Spanish has been strengthening primary categorization at the lowest levels.

¹⁰ Mandarin root categoriality is not so low in this field. Bisang (2006) shows that root categoriality is extremely low in Late Archaic Chinese. It seems plausible that root categoriality has been rising in the history of Chinese. Cf. fn. 4.

¹¹ Lest anybody think that uncategorized roots are created by an artifact of the analysis, viz. by truncating stems, it should be repeated (cf. § 3.2) that many roots do accept alternate thematic vowels or other derivational suffixes. For instance, from the stem *laeto-* ‘merry’, we obtain the root *laet-* by subtracting the thematic vowel. Although this root is no stem by itself, it also serves as the basis for the verb stem *laet-a-* ‘rejoice’, with a different thematic vowel. Similarly, by subtracting the thematic vowel from the stem *ama-* ‘love’, we obtain the root *am-*, which also serves as the basis for the noun stem *am-or* ‘love’, with a different derivational suffix. By the above-mentioned principle ‘once a morpheme – always a morpheme’, the same operation is applied to stems such as *ir-a* ‘wrath’, although the root *ir-* is otherwise useless.

3.4.3 Categoriality of roots and stems

If we only look at the languages occupying the poles of the two scales of stem categoriality (T7) and root categoriality (T8), there appears to be no connection between the two scales. This changes if we arrange them side by side, as in T9:

T9. *Root and stem categoriality*

language	root	stem	difference
Latin	0.30	0.99	0.69
Spanish	0.59	0.98	0.29
Mandarin	0.60	0.84	0.24
Yucatec	0.76	0.96	0.20
English	0.62	0.77	0.15
German	0.78	0.86	0.08
average	0.61	0.90	0.28

Here it is at once evident that, for each language in the sample, stem categoriality is higher than root categoriality. This is the basic generalization to be made despite the enormous range of variation represented in the last column. This time, the poles are occupied by Latin and German. Latin has productive processes of stem formation, and it uses them essentially in order to categorize roots. German, on the other hand, has no such processes, nor does it need them, because its roots are largely pre-categorized.

While the universal principle is clear, the question remains whether the enormous cross-linguistic differences are typologically relevant. Does the difference between root categoriality and stem categoriality correlate with anything else? Here I can only hint at a factor that appears to play a role, the morphological complexity of stems. T10 shows the languages again arranged according to the difference in categoriality between roots and stems and confronts these values with the ratio of morphs per stem:

T10. *Categoriality difference and morphs per stem*

value language	categoriality difference	morphs per stem
Latin	0.69	2.02
Spanish	0.29	1.54
Mandarin	0.24	1.82
Yucatec	0.20	1.45
English	0.15	1.29
German	0.08	1.27
average	0.28	1.57

It appears there is a flawless correlation between the two values, with the one exception of Mandarin. Now, the exceptional behavior of Mandarin is instructive. Many derivational operators just as stem-forming operators are structural heads, bestowing their category onto the composite whole. In Mandarin, however, complex stems are not formed by such operators, but by compounding, where the structural head of the compound is often less than clear. Here one root is just like the other, i.e. if one root does not suffice to determine the target category, then compounding it with a second root will not help very much. We therefore tabulate the results once more, this time considering not the ratio of morphs per stem, but the ratio of non-root morphs per stem. The result is shown in T11.

T11. *Categoriality difference and non-roots per stem*

values language	categoriality difference	non-roots per stem
Latin	0.69	0.88
Spanish	0.29	0.54
Mandarin	0.24	0.0
Yucatec	0.20	0.32
English	0.15	0.28
German	0.08	0.23
average	0.28	0.375

Again, the two values correlate nicely, and again Mandarin is the only exception. Since the lexemes of the Mandarin sample are composed exclusively of roots, the ratio of non-roots per stem is zero. Despite the absence in Mandarin of operators that could confer a target category onto a base in a regular way, categoriality increases in stems over roots by 0.24, which is almost the cross-linguistic average. In a loose way of speaking, we may say that the increase in categoriality from roots to stems is, in principle, brought about by operators of stem formation, most efficiently by Latin-type

thematic formatives. Failing that, sheer morphological complexity raises categoriality, anyway, though less efficiently.¹²

4. Conclusion

The piece of research reported here suffers from lack of methodological rigor in various respects:

- the empirical domain, i.e. the set of stems to be investigated, has to be delimited more exactly,
- decisions concerning inclusion and exclusion of data must be more principled,
- cross-linguistic comparability of data must be secured more systematically,
- morphological analysis must be refined,
- categoriality must be measured in a more formal way.

Future research with better methodology will doubtless modify the results obtained here. It will also be necessary to extend the research to more languages. Another look at T7 reveals that the sample, small and biased as it is, contains a language (Latin) that occupies the pole of extreme stem categoriality. On the other hand, the lowest stem categoriality in the sample is 0.77 (English). There are doubtless languages closer to the zero pole. A candidate is Samoan. According to Mosel & Hovdhaugen 1992:73-77, all Samoan words can be used as heads of noun, verb or modifier phrases. As for root categoriality (T8), Samoan would again occupy the zero pole.¹³ Neither does the sample contain a language close to the pole of high root categoriality (the highest value in T8 being 0.78 for German). Here Persian may be a candidate.¹⁴

Pending such methodological refinements and expansions of the database, I assume that the generalization concerning **staggered level-dependent categoriality** of linguistic signs is likely to survive. For the two lowest levels of T1, the stem and the root, the present research has proved that categoriality is consistently higher at the stem level, thus, at the higher level.¹⁵ And we know independently that every syntagm has a unique category at the sentence level. We can therefore safely generalize that **categoriality increases with the grammatical levels**.¹⁶ This makes sense in view of our initial assumption (§ 2) that if anywhere, then

¹² The correlation between categoriality difference and morphological complexity is trivial in two precise cases: In a perfectly isolating language, root equals stem. Here the categoriality difference and morphological complexity are both zero. No language in the world comes even close to this ideal. The most isolating languages, like Archaic Chinese and Vietnamese, are heavily compounding, similarly to Mandarin in my sample. Here stems are morphologically more complex than roots, and the categoriality increases, too. The other case where the correlation is trivial is a language where stems differ from roots by a stem-forming operator. Latin does come close to this type (cf. Haig 2006:6). Even in such a language, however, there is no logical necessity for root categoriality to be lower than stem categoriality.

¹³ Mosel & Hovdhaugen (1992:77) sometimes refer to “words”, sometimes to “roots”. If what they say is true for words, then by the principle of § 3.4.3 above, it would have to hold *a fortiori* for stems.

¹⁴ Avazeh Mache (University of Erfurt) translated the concept list into Persian and found that 100% of the roots were uniquely categorized.

¹⁵ This fits nicely with Hopper & Thompson’s (1984) and Croft’s (1991:48) finding that categoriality of a stem decreases if it becomes a component of another stem.

¹⁶ A similar claim is made in Haig 2006:49: “If precategoriality is a feature anywhere in a grammar, then in its deepest levels.” He suggests a distinction between “early and late categorizing languages” with respect to a passage through T1 from bottom to top. In Hawkins 2007, it is hypothesized that if a language does not do the categorization at the lexical level, it will need more categorizing apparatus at the syntactic level.

structural categories are needed at the level of the sentence.¹⁷ Therefore, a necessary extension of the research started here would include an analysis of categoriality at higher levels.

Diachronically, roots are lexicalized stems. Engl. *thirst* is originally a derived abstract noun to a Germanic verb meaning ‘to dry’, and similarly German *Angst* is originally derived from a verb etymologically identical with the Latin verb *angere* ‘frighten’, which is in the sample, too. In a certain perspective, lexicalization (like grammaticalization) is a reduction process where information is lost. If a stem is lexicalized to a root, that loss may affect different kinds of features. If categorial information is lost, then lexicalization is the diachronic manifestation of the categoriality difference between stems and roots that was ascertained in § 3.4.3 for each language at the synchronic level. If, however, no categorial information is lost in lexicalization, then the root that is output of the process may inherit the category of the stem that is its input. If that happens on a large scale and if roots had low categoriality at the input stage, then the result of the change will be an increase in overall root categoriality. This is apparently what happened on the way from Latin to Spanish.

There have been “theories” in the recent literature which claim that universally roots are categorially indeterminate in the lexicon and that it is the syntactic context that determines their category. Thus, Hopper & Thompson (1984:747) argue in a functional-typological framework “that linguistic forms are in principle to be considered as LACKING CATEGORIALITY completely unless nounhood or verbhood is forced on them by their discourse functions.” In a variant of “distributed morphology”, Marantz (1997) denies categorial status to roots. Such theories suffer from a methodological and a theoretical misconception. The methodological mistake is that they try to resolve at the theoretical level what is a purely empirical issue. Eugenio Coseriu once wrote (1958:109):

La idea de juntar hechos para resolver los problemas teóricos es una idea reaccionaria que implica detener la investigación y no fundarla más sólidamente, como se pretende; es, en los casos extremos, una forma típica de misologismo que pretende presentarse como cautela científica.¹⁸

There is nothing one could reasonably object to this. There is just one thing that should be added:

The idea to postulate a set of principles in order to solve empirical problems is an escapist idea which only serves to delay research and not – as is often claimed – to provide a more solid foundation; in the extreme cases, it is a typical form of arm-chair linguistics posing as theoretical linguistics.

Sometimes one feels tempted to remind those “theorists” of Martin Joos’s famous dictum “that languages could differ from each other without limit and in unpredictable ways” (Joos (ed.) 1957:96); not because Joos was literally right, but because he formulates the appropriate

¹⁷ Another important theoretical issue that I am not resolving here concerns the model character of one level of T1 for categorizations made at another level. The above would invite the inference that higher levels dictate the necessary categories, and categorizations made at lower levels follow that model. On the other hand, there is Dik’s (1985) Principle of Formal Adjustment of Derived Constructions, according to which derived constructions are coined on the model of basic constructions. That would seem to entail that categories of lower levels serve as models for categorization at higher levels.

¹⁸ The idea to accumulate facts in order to solve theoretical problems is a reactionary idea which only serves to delay research and not – as is often claimed – to provide a more solid foundation; in the extreme cases, it is a typical form of averseness to logic posing as scientific caution.

methodological attitude to empirical issues. Whether or not roots are category-specific is such an empirical question. The methodologically sound position is to be prepared for cross-linguistic variation in every respect that is not of logical necessity invariable.¹⁹

The theoretical mistake consists in positing universal properties of categories of grammar, in this case the precategoriality of roots. Languages are problem-solving systems. Some of the tasks to be solved are universal. The task of assigning expressions to categories is among these, so it must be incorporated into linguistic theory. However, incorporating one of the possible solutions into linguistic theory fails to recognize that the solution of a problem is dynamic in nature and there are generally alternative ways of solving a problem. The task of universals research is not to stipulate one of the possible solutions as a property of universal grammar, but to analyze the variation encountered in order to identify its principle.

One important result of this empirical investigation is that there are languages, in particular German, with a strong primary categorization of roots.²⁰ If we had investigated a lexical field such as fruit trees, then a high categoriality of members of the field – most of them beings nouns – would be a rather trivial result of the cognitive principle of grammatical categorization mentioned in § 2. However, the data adduced are from the field of experience. There are no known cognitive principles that would regulate the grammatical categorization of roots meaning ‘calm’ or ‘dread’, and indeed there is wide variation to be found:

- both inside a language and cross-linguistically, roots designating such concepts belong to a wide range of categories,
- and in every language, such roots have relatively low categoriality.

In view of this, it is all the more remarkable that there are languages that apply a primary categorization to most of their roots although it is not so clear what it is needed for. Theoretically, it could suffice to have precategorial roots, to categorize them once at the level of the stem and then assign them their final category at the level of the sentence. Some languages like Latin come close to that model. It is no coincidence that the modists came up with a theory of precategorial roots, since their linguistics was exclusively based on Latin. However, some languages do it differently. Some like German do the bulk of the categorization at the level of the root, with stem formation adding little to that. As we have seen, categorization at the level of the stem is essentially the result of morphological operations. Again, categorization at the syntactic level essentially means inserting the item in a syntactic template. All of this requires structural apparatus; it raises structural complexity. The German position is therefore: Concerning categorization of signs, do at the root-level what can be done at the root-level, and reserve structural complexity for other functions.

Taking up a theoretical consideration introduced in § 2: The *raison d'être* of a sentence is to serve as an utterance. More generally, the *raison d'être* of (virtual) *langue* is to render (actual) *parole* possible. Consequently, the categorization of units of *langue*, like roots and stems, is, so to speak, a preliminary categorization or precategorization whose purpose is to unburden categorization in *parole*.

On the basis of the data gathered, other kinds of questions may be approached. For instance: For each language, which category prevails in the roots of each of the subdomains? Are there universal

¹⁹ Cf. Lombardi Vallauri 2004 for this position in general, and Haig 2006:41 for precategoriality as a typological parameter rather than a universal, in particular.

²⁰ Don 2004 argues the same for Dutch.

tendencies in this respect; are there typologically relevant differences among the languages?²¹ However, answers to such questions presuppose that first the methodology be refined as required above.

²¹ Cf. Verhoeven 2007, ch. 5.4.1 for the categorial profile of Yucatec Maya in the domain of experience.

Appendix

One example from each of the four subdomains in the six languages.

language	stem	categories	categoriality	root	categories	categoriality
English	<i>hunger</i>	N/V	0.5	<i>hunger</i>	N/V	0.5
German	<i>Hunger</i>	N/V	0.5	<i>hunger</i>	N/V	0.5
Latin	<i>fame-s</i>	N	1	<i>fam-</i>	X	0
Spanish	<i>hambre</i>	N	1	<i>hambr-</i>	X	0
Mandarin	<i>è</i>	V	1	<i>è</i>	V	1
Yucatec	<i>wi'h</i>	A	1	<i>wi'h</i>	A	1
English	<i>excite</i>	V	1	<i>cite</i>	X	0
German	<i>aufreg-</i>	V	1	<i>reg-</i>	V	1
Latin	<i>concita-</i>	V	1	<i>cit-</i>	X	0
				<i>con-</i>	Adv	1
Spanish	<i>agit</i>	V	1	<i>agit</i>	V	1
Mandarin	<i>jīdòng</i>	V	1	<i>jī</i>	V	1
				<i>dòng</i>	V	1
Yucatec	<i>péek'ool</i>	V	1	<i>peek</i>	V	1
				<i>ool</i>	N	1
English	<i>greed</i>	N	1	<i>greed</i>	N	1
German	<i>Gier</i>	N/V	0.5	<i>Gier</i>	N/V	0.5
Latin	<i>ave</i>	V	1	<i>av</i>	X	0
Spanish	<i>afán</i>	N	1	<i>afán</i>	N	1
Mandarin	<i>tānkán</i>	N/V	0.5	<i>tān</i>	V	1
				<i>kán</i>	X	0
Yucatec	-					
English	<i>imagine</i>	V	1	<i>imagine</i>	V	1
German	<i>vorstell-</i>	V	1	<i>stell</i>	V	1
Latin	<i>cogita</i>	V	1	<i>con</i>	Adv	1
				<i>ag</i>	V	1
Spanish	<i>imagin-</i>	V	1	<i>imagin</i>	V	1
Mandarin	<i>xiǎngxiàng</i>	N/V	0.5	<i>xiǎng</i>	V	1
				<i>xiàng</i>	N/V	0.5
Yucatec	<i>wayáak'</i>	N/V	0.5	<i>wayáak'</i>	N/V	0.5

References

- Bisang, Walter 2006, "Transcategoriality and syntax-based parts of speech - the case of Late Archaic Chinese. Paper read at 'Universality and Particularity in Parts-of-Speech Systems', Amsterdam, 8-10 June 2006. Mainz: Universität Mainz.
- Bursill-Hall, G. L. 1972. *Grammatica Speculativa of Thomas of Erfurt*. London: Longman.
- Coseriu, Eugenio 1955, "Sobre las categorías verbales ("partes de la oración")." *Revista de Lingüística Aplicada* 10:7-25. Repr.: Coseriu, Eugenio 1978, *Gramática, semántica, universales. Estudios de lingüística funcional*. Madrid: Gredos (Biblioteca Románica Hispánica, II/280); 50-79.
- Coseriu, Eugenio 1958, *Sincronía, diacronía e historia. El problema del cambio lingüístico*. Montevideo: Universidad de la República de Uruguay, Facultad de Humanidades y Ciencias (2. ed., rev. y corr.: Madrid: Gredos (Biblioteca románica hispánica, 2,193), 1973.
- Croft, William 1991, *Syntactic categories and grammatical relations. The cognitive organization of information*. Chicago: Chicago University Press.
- Dik, Simon C. 1985, "Formal and semantic adjustment of derived constructions." Bolkestein, A. Machtelt *et al.* (eds.), *Predicates and terms in functional grammar*. Dordrecht & Cinnaminson: Foris (Functional Grammar Series, 2); 1-28.
- Don, Jan 2004, "Categories in the lexicon." *Linguistics* 42:931-956.
- Haig, Geoffrey 2006, "Word-class distinctions and morphological type: agglutinating and fusional languages reconsidered." Kiel: Seminar für Sprachwissenschaft der Universität (Unpubl. ms.)
- Haspelmath, Martin 1993, "More on the typology of inchoative/causative verb alternations." Comrie, Bernard & Polinsky, Maria (eds.), *Causatives and transitivity*. Amsterdam & Philadelphia: J. Benjamins (Studies in Language Companion Series, 23); 87-120.
- Hawkins, John A. 2007, "Nouns and noun phrases: grammatical variation and language processing." Paper read at the Symposium 'Nouns cross-linguistically', 22-23 June 2007, Università degli Studi del Molise, Campobasso.
- Hopper, Paul J. & Thompson, Sandra A. 1984, "The discourse basis for lexical categories in universal grammar." *Language* 60:703-752.
- Joos, Martin (ed.) 1957, *Readings in linguistics. The development of descriptive linguistics in America since 1925*. New York: American Council of Learned Societies.
- Kuryłowicz, Jerzy 1936, "Dérivation lexicale et dérivation syntaxique (contribution à la théorie des parties du discours)." *Bulletin de la Société de Linguistique de Paris* 37:79-92.
- Lehmann, Christian 2005, "Zum Tertium Comparationis im Sprachvergleich." Schmitt, Christian & Wotjak, Barbara (eds.), *Beiträge zum romanisch-deutschen und innerromanischen Sprachvergleich. Akten der gleichnamigen internationalen Arbeitstagung (Leipzig, 4.10.-6.10.2003. 2 Bde. Bonn: Romanistischer Verlag; 1:157-168.*
- Lehmann, Christian & Shin, Yong-Min & Verhoeven, Elisabeth 2000, *Person prominence and relation prominence. On the typology of syntactic relations with special reference to Yucatec Maya*. München: LINCOM Europa (LINCOM Studies in Theoretical Linguistics, 17).
- Lois, Ximena & Vapnarsky, Valentina 2003, *Polyvalence of root classes in Yukatekan Mayan languages*. München & Newcastle: LINCOM Europa (LINCOM Studies in Native American Linguistics, 47).
- Lombardi Vallauri, Edoardo 2004, "The relation between mind and language: The Innateness Hypothesis and the Poverty of the Stimulus." *The Linguistic Review* 21:345-387.
- Marantz, Alec 1997, "No escape from syntax. Don't try morphological analysis in the privacy of your own lexicon." Dimitriadis, Alexis *et al.* (eds.), *Proceedings of the 21st Annual Penn Linguistics Colloquium*. Philadelphia: University of Pennsylvania Press (University of Pennsylvania Working Papers in Linguistics, 4/2); 201-225.
- Misteli, Franz 1893, *Charakteristik der hauptsächlichsten Typen des Sprachbaues. Neubearbeitung des Werkes von Prof. H. Steinthal (1861)*. Berlin: F. Dümmler (Steinthal, Heymann & Misteli, Franz, *Abriß der Sprachwissenschaft*, Bd. II).

- Mosel, Ulrike & Hovdhaugen, Even 1992, *Samoan reference grammar*. Oslo: Norwegian University Press.
- Simon, Walter 1937, "Has the Chinese language parts of speech?" *Transactions of the Philological Society* 99-119.
- Verhoeven, Elisabeth 2007, *Experiential constructions in Yucatec Maya. A typologically based analysis of a functional domain in a Mayan language*. Amsterdam & Philadelphia: J. Benjamins (Studies in Language Companion Series, 88).