Lund University, Dept. of Linguistics and Phonetics
Working Papers 53 (2008), 5–14

Phonological erosion and semantic generalization: Notes on the grammaticalization of the Tocharian case paradigm[1]

Gerd Carling

0 Background

In Carling 2000 I investigated the local cases of Tocharian denoting motion and location, namely the locative, perlative, allative, and oblique of direction. The last chapter, *Rekonstruktion der Vorgeschichte des lokalen Paradigmas,* dealt with the formal reconstruction of the case paradigm, from which I tried to establish a probable functional evolution of the respective cases. In Carling 1999 I dealt more comprehensively with the morphological reconstruction and the gradual development of the case paradigm of pre-Tocharian.

In this article I will try to add some theoretical dimensions to the arguments presented in Carling 1999, 2000. I will focus on the difference between the surface-oriented linguistic alteration, i.e. the phonological and morphological grammaticalization, and the linguistic change on the functional/semantic level.

1 Some theoretical prerequisites

Considering the evolution of a case system like the one reconstructed for pre-Tocharian, we can expect linguistic change to have taken place on four levels, as described by Lehmann 1985:108; cf. also Dik 1997:49:

(a) The level of communicative sentence perspective, where the opposition between *theme* vs. *rheme* and *topic* vs. *focus* are the most important.

[1]I thank dr. Judith Josephson, Göteborg, for correcting my English.

(b) The level of sentence semantics, where we have semantic roles, such as agent, patient, etc.

(c) The level of syntax, where we have syntactic functions such as subject and direct object, absolutive or ergative, etc.

(d) The level of morphology, where we have *cases*, such as nominative and accusative, absolutive and ergative, etc.

Since Tocharian, as we will observe, is a language the prehistory of which can only be reconstructed, not observed, it is important to take into consideration the plausibility of a reconstruction. According to Givón 1999:95, 2000:12, "proposed diachronic changes must conform to what is known about universals of diachronic change", of which there are three kinds: *semantic, phonological* and *typological* plausibility.

The levels of primary interest for a reconstruction of grammaticalization of a case paradigm are (b), (c) and (d). The level of morphology (d) can be reconstructed with a relatively large degree of certainty. The reconstruction of syntax and semantics rests mainly upon the reconstruction of morphology. Reconstruction of communicative sentence perspective is of less interest when dealing with paradigmatic innovation, which has been noticed by Givón 2000:12): "Grammatical and morphological innovation tends to occur in the most common, neutral clause type (main, declarative, affirmative, active)".

2 The evolution of the Tocharian paradigm

For Tocharian A and B, only one linguistic stage is attested. The chronological range of approximately four or five hundred years, which our text material covers, does not provide enough material to establish more than smaller linguistic changes. What we have are two quite well established linguistic units, Tocharian A and Tocharian B. However, a comparison of these two languages enables us to reconstruct a Proto-language, Common Tocharian, and opens a wider perspective for the prehistory of Tocharian A and Tocharian B, respectively. Otherwise, changes prior to our attested paradigms have to be looked for through the glasses of internal reconstruction.

In Carling 1999:98 I suggest four reconstructed stages of the pre-Tocharian case paradigm, which follow upon an initial break-down of a presumed Indo-European eight-case system:

Stage I: a supposed system containing only the primary (inflectional) cases, before the building up of an agglutinative paradigm; case functions

must have been expressed analytically. At this stage we have a paradigm of three cases of Indo-European origin, all of them preferably grammatical or *core cases*: nominative, oblique and a merged genitive/dative case. Languages with only grammatical cases or, to be precise, cases of which the main function is grammatical rather than semantic, are well attested (cf. Lehmann 1983:368).

Stage II: the reconstructed Common Tocharian paradigm, containing the primary cases, and the secondary cases locative, allative and perlative. This paradigm is an earlier variant of the later paradigm of Tocharian A and Tocharian B. The difference between the primary and the secondary cases as representing *core* vs. *peripheral cases* must have been quite evident. The secondary cases were fundamentally local cases, denoting location (locative), direction *towards* (allative), and motion *along* (perlative). These remained their basic functions in Tocharian A as well as Tocharian B (more about this in section 4).

Stage III: the Common Tocharian paradigm with the addition of the post-Common Tocharian secondary affixes ablative, comitative, A instrumental and B causal = pre-A and pre-B. This is the system of Common Tocharian, extended by more cases, but using the same agglutinative principle. The case markers (affixes) of these above mentioned cases are different in Tocharian A and Tocharian B, and they are evidently formed by adpositional elements.

Stage IV: Tocharian A and Tocharian B. Here we find our two attested paradigms with nine cases each. In relation to the pre-Tocharian paradigm(s), the case affixes have been affected by linguistic change, and the functions have expanded semantically (more about this in sections 3 and 4).

It is, of course, impossible to know whether the process of forming new cases was finished, or if Tocharian A and B would have continued increasing their systems if the languages had not become extinct. As far as our two paradigms are concerned, there was a clear tendency toward decreasing or cliticization of the case affixes, which was more developed in Tocharian A than in Tocharian B. This indicates that the system could have become reduced, if the languages had continued.

3 Effects of grammaticalization on the pre-Tocharian paradigm

3.1 Reanalysis and analogy

Reanalysis and *analogy* could be designated as the triggering linguistic mechanisms behind grammaticalization. Reanalysis and analogy are separate

processes which operate on different levels of the language, as is very precisely described by Hopper & Traugott 1993:61:

> Reanalysis essentially involves linear, syntagmatic, often local reorganization and rule change. It is not directly observable. On the other hand, analogy essentially involves paradigmatic organization, change in surface collocations, and patterns of use. Analogy makes the unobservable changes of reanalysis observable.

Both these processes can be exemplified by the Tocharian paradigm. However, since analogy is the most surface-oriented of these processes, it can easily be established through comparative morphology; reanalysis, on the other hand, must be analyzed basically through the effects of analogy.

As an example of analogy we may consider the resegmentation of the affixes of the Common Tocharian paradigm (cf. Pinault 1989:74f., Carling 1999:96):

all.sg.	*yäkwæ	-cä	loc.sg.	*yäkwæ	-næ
	horse-OBL	-ALL		horse-OBL	-LOC
all.pl.	*yäkwæ-ns	-cä	loc.pl.	*yäkwæ-ns	-næ
	horse-OBL.PL	-ALL		horse-OBL.PL	-LOC

The general structure for the formation of cases was agglutinative, which can be schematized as follows:

[NOUN-OBL(SG/PL)-CASE]

Thereupon, a generalization took place in Tocharian A as well as Tocharian B in which the thematic vowel in Tocharian A became part of the affixes, whereas in B the oblique plural ending *-*s* became part of the affixes, as follows:

Pre-A:	all.sg.	[*yäkwæ#cä]	>	[*yäkw#æcä]	> A [yuk#ac]
	loc.sg.	[*yäkwæ#næ]	>	[*yäkw#ænæ]	> A [yuk#aṃ]
Pre-B:	all.pl.	[*yäkwæ#ns#cä]	>	[*yäkwæ#n#scä]	> B [yakwe#ṃ#ś(c)]
	loc.pl.	[*yäkwæ#ns#næ]	>	[*yäkwæ#n#snæ]	> B [yakwe#ṃ#ne]

This yielded new case affixes beginning with *-a-* in Tocharian A, where the singular became the starting point for an analogical generalization. In Tocharian B, we find a different situation, yielding new affixes beginning with *-s-*. Here, the plural became the starting point for an analogical

generalization. Still, we have the same structure for the formation of cases as we had in Common Tocharian:

[NOUN-OBL(SG/PL)-CASE]

but with the 'new' affixes described above.[2]

We should suspect reanalysis as the triggering factor behind analogical generalizations as those observed in the pre-Tocharian paradigm. Since we can reconstruct an adpositional origin for most of our Common Tocharian affixes, we can postulate an earlier variant:

[[NOUN-OBL] [ADP]]

that was reanalysed as:

[NOUN-OBL-CASE]

3.2. Phonological erosion

At the level of morphology, the cases were typically affected by *phonological erosion* or *phonological reduction* (for a general overview, see Bybee, Perkins & Pagliuca 1994:6f.). We can observe the following effects of this procedure (for the definitions, cf. Lessau 1994:260):

(a) *Peripheral erosion* resulted in the loss of final syllables, as e.g. in pre-A for the ablative: **-æ-ṣu* (from A *ṣu* prev. 'away from') > **-æṣ* (cf. above) > A *-äṣ*, or for the comitative: **-æ-śälæ* (from **śälæ* adp. 'together with', A *śla* B *śale, śle*) > **-æśäl* (cf. above) > A *-aśśäl*.

(b) *Junctural erosion* resulted in loss of phonemes at morpheme boundaries, which occured e.g. in the locative plural (which was then generalized to the singular) in pre-B: **-ns-næ* > **-n-snæ* (cf. above) > **-ṃ-næ* > B *-ṃ-ne*.

3.3. Semantic bleaching

As concerns the morphosemantic side, the original adpositions or particles, used as case affixes, were typically affected by *semantic bleaching*, *desemanticization* or *semantic generalization,* as e.g. defined by Heine & Reh 1984:36: "... a lexical item receives a second, non-lexical function, which may ultimately become its only function.". This means that the lexical

[2]A process similar to this has been described for other languages, for example Samoyedic (Mikola 1975:170-2).

content of the original adpositions, becoming case affixes, was successively lost and was replaced by or reduced to a more or less grammatical content.

As noticed by several authors (cf. Hopper & Traugott 1993:87), this evolution is twofold: On the one hand, a more or less specialized semantic content and a narrowed syntactic use become generalized. This is typically the pattern for an adposition becoming a case marker, as for instance in the case of A *ṣu* 'away (from)', a preverb not very frequently used, and the ablative suffix *-äṣ*, which has a wider range of uses. On the other hand, this process results in a shift in semantic content, and a new meaning is gained in the process, which has been described in terms of 'pragmatic enrichment' (Hopper & Traugott 1993:87). Bybee, Perkins & Pagliuca 1994:289 describe this process very precisely as follows: "generalization is the loss of specific features of meaning with the consequent expansion of appropriate contexts of use for a gram".

The process outlined above should be designated as the first grammaticalization process in which adpositional elements become case markers. This process could be subsumed under the general notion *cliticization* (Givón 2000:121). For Tocharian, this process can only be observed indirectly, since it takes place during reconstructed stages of pre-Tocharian. The second process, which will be dealt with in the next part, is the grammaticalization of the case functions that started to be in operation when the initial process, described above, was in the process of completion.

4 (Channels of) semantic generalization

As I demonstrated in Carling 2000:384ff., there is a clear tendency for the 'older' cases, i.e. locative, perlative and allative, to have a wide range of functions, whereas the later formed cases, ablative, instrumental, comitative and causal are much more restricted in use. In Carling 2000 I proposed that the Common Tocharian cases locative, perlative and allative were originally simple local cases, denoting *location, motion along* and *direction towards*. Thereupon, the cases were successively affected by an expansion of semantic content and syntactic use. This means that in the case of the Tocharian locative, perlative and allative, that their function, as well as their syntactic use, was expanded to cover other functional areas, different from their original ones.

In terms of grammaticalization, this expansion started from the lowest, most concrete level, i.e. with a simple local function (cf. below), and expanded higher up in the case hierarchy. It is important to note, however,

that this change should not be thought of in terms of an 'evolution', where one function becomes another function and the first function is lost, as we defined semantic bleaching in the previous section. Rather, this is a true *expansion*, in which the original function of the case is kept, but the functional domain linked to the case is widened. Here it is important to note that, in spite of gaining grammatical functions and syntactic uses, the original concrete, local function remained the basic function, as we will see later.

But how do we define a case function as being more grammaticalized than another? At the top we find, as expected, the cases expressing the three core functions S (Subject), A (Agent) and P (Patient). On the lowest, least grammatical level most authors tend to put concrete, local functions (Heine, Claudi & Hünnemeyer 1991:156, Blake 1994:89, Dik 1989:226). Otherwise, there are different views as to what extent the parameters [±abstraction] and [±animacy] change the degree of grammaticalization of a function. There seem to be different views of the degree of grammaticalization of the function TIME.[3] In Tocharian, most of our secondary cases (perlative, locative, instrumental but not the allative) as well as the oblique, are used in temporal constructions. I suspect, with Heine, Claudi & Hünnemeyer 1991:151, who have schematized the development of the ALLATIVE case marker in Ik and Kanuri, that the development of temporal functions formed a separate line in the grammaticalization of the local cases.[4]

If we consider the allative, its original function was to denote direction *towards*. This function did not change, but remained the basic function of the case on into Tocharian A and Tocharian B. Second, we find that it has the expanded function of expressing first argument of the verbs 'look (at), behold' and 'tell' (PATIENT) as well as 'flatter', 'trust' (BENEFICIARY). This function, compared to the former, more concrete local functions, represents a more abstract level and thus can be seen as more grammaticalized. Further, the allative is used to indicate the indirect object with certain verbs, such as 'show', 'send', 'present with' (BENEFICIARY); in this function the allative competes with the genitive, which is the main case for marking indirect object (and is also the morphological correspondant of both the Indo-European dative and genitive). Further, the allative developed the function of PURPOSE in a limited number of circumstances, but it never developed any temporal uses.

[3]Cf. Dik 1989:74f. who puts TIME on the lowest level and Heine, Claudi & Hünnemeyer 1991:159, who put TIME on a very high level.
[4]For fuller information to the following review of the case functions I refer to Carling 2000:5ff.

The perlative displays a more complicated pattern. The local function of the perlative was basically to denote motion *along* or *over* (PERLATIVE), and this function remained very important in Tocharian A and Tocharian B. This was, however, transmitted to the notion of 'beside' or 'over', independent of motion or location (ADESSIVE etc.), as opposed to the locative, which denoted motion *into* or location *in* (INESSIVE, ILLATIVE). It is quite evident that the core meaning of the perlative was local, but *exactly* which kind of local function is more difficult to ascertain.

Other, more abstract, but still peripheral functions of the perlative were to denote MANNER or CAUSE. At a more grammaticalized level the functions of INSTRUMENTAL and AGENT (in passive constructions) emerged. In Tocharian A a new case, instrumental, was formed to denote INSTRUMENTAL. This case was also grammaticalized and could be used in the role of AGENT. This resulted in a situation in Tocharian A, in which instrumental was used as AGENT with non-animate objects and perlative with animate objects. Considering the animacy hierarchy (cf. Dik 1989:32): HUMAN > ANIMATE > INANIMATE [+force] > INANIMATE [–force], the perlative appears as more grammaticalized than the instrumental in this function.

The locative was basically a case denoting location, which developed into the function of marking location *in* or motion *into*, as opposed to the perlative. The locative developed several abstract, but still peripheral functions, but it never reached the degree of grammaticalization of the perlative and the allative. Among functions on the borderline to grammatical usage one may mention that it was used as complement to the verbs 'be angry (with)' or 'be attached to'.

Lehmann 1985:128 describes the notion of *desemanticization* as follows:

> ... at the source of a grammaticalization process, we have *Grundbedeutung*, (core meaning) of the item; at the end, we have its *Gesamtbedeutung* (general meaning). This relationship manifests itself both diachronically, as the semantic gradation ... and synchronically, as a specific kind of polysemy.

This description fits the second desemanticization process of Tocharian very well: the cases kept a solid basic function that remained in use even though the semantic area was expanded. It is important to note that this desemanticization is different from that usually described as following cliticization (i.e. the process of a lexeme becoming a morpheme), where the

original semantic content of the lexeme becomes bleached and is finally replaced.

5 Summary

To sum up, we can observe a manifold evolution of the case system which operated on different levels of the language. At first, the system was greatly delimited: the Indo-European case system was stripped of all its peripheral cases, leaving only a minimized paradigm of core cases. Thereupon, a new system was successively re-built: through analysis new cases were formed from adpositional elements, which resulted in cliticization of these items: phonological erosion and analogical levelling affected the shape of the items, and bleaching changed their semantic content. While this process was in the process of completion, a second wave of desemanticization started to operate on the function and syntactic use of the case markers. They expanded their functional content, and gained new, more abstract and more grammatical functions, which in turn lead to a wider syntactic use of the cases. They retained, however, the original concrete, local function, as their basic function.

References

Blake, Barry J. 1994. *Case*. Cambridge: Cambridge University Press.

Bybee, Joan, Revere Perkins & William Pagliuca. 1994. *The evolution of grammar: tense, aspect, and modality in the languages of the world*. Chicago: University of Chicago Press.

Carling, Gerd. 1999. 'The Tocharian inflected adverbials and adpositions in relation to the case system', *TIES* 8, 96-110.

Carling, Gerd. 2000. *Die Funktionen der lokalen Kasus im Tocharischen*. Berlin: Mouton de Gruyter.

Dik, Simon. 1989. *The theory of functional grammar*. Part I: The structure of the clause. Dordrecht: Foris Publications.

Dik, Simon. 1997. *Theory of functional grammar*. Second, revised edition, ed. by Kees Hengeveld (in two volumes). Berlin: Mouton de Gruyter.

Givón, Talmy. 1999. 'Internal reconstruction, on its own'. In Edgar C. Polomé & Carol F. Justus (eds.), *Language change and typological variation: in honor of Winfred P. Lehmann on the occasion of his 83rd birthday*, volume I: Language change and phonology, 86-130. Washington: Institute for the Study of Man.

Givón, Talmy. 2000. 'Internal reconstruction: as method, as theory'. In Spike Gildea (ed.), *Reconstructing grammar: comparative linguistics and grammaticalization*, 107-159. Amsterdam: John Benjamins.

Heine, Bernd, Ulrike Claudi & Friederike Hünnemeyer. 1991. *Grammaticalization: a conceptual framework*. Chicago: University of Chicago Press.

Heine, Bernd & Mechthild Reh. 1984. *Grammaticalization and reanalysis in African languages*. Hamburg: Helmut Buske Verlag.

Hopper, Paul J. & Elizabeth Closs Traugott. 1993. *Grammaticalization*. Cambridge: Cambridge University Press.

Lehmann, Christian. 1983. 'Rektion und syntaktische Relationen'. *Folia Linguistica* 17, 339-378.

Lehmann, Christian. 1985. *Thoughts on grammaticalization*. München: Lincom Europa.

Lessau, Donald (ed.) 1994. *A Dictionary of grammaticalization*. 3 vols. Bochum: Universitätsverlag Dr. N. Brockmeyer.

Mikola, Tibor. 1975. *Die alten Postpositionen des Nenzischen (Jurak-samojedischen)*. Den Haag: Mouton.

Pinault, Georges-Jean 1989. 'Tokharien'. In *LALIES* 7, *Actes des sessions de linguistique et de littérature*. Paris: Presses de l'École Normale Supérieure, 6-224.

Lund University, Dept. of Linguistics and Phonetics
Working Papers 53 (2008), 15–21

Why is the Good distribution so good? Towards an explanation of word length regularity

Mats Eeg-Olofsson

Abstract
In 2004, Sigurd, Eeg-Olofsson & van de Weijer fitted the discrete analogue of the statistical gamma distribution to the frequency of the length of various linguistic units, in particular the length in letters of the word tokens in English and Swedish corpora. This distribution is also known as the Good distribution (Johnson, Kemp & Kotz 2005), named after I. J. Good, the statistician. In 2005, Lupsa & Lupsa successfully fitted this distribution also to the length of the base forms in Romanian and English dictionaries. Without further motivation, Lupsa & Lupsa call this regularity a linguistic law. This paper presents data from various languages to show that it is indeed a candidate for a linguistic universal and hints at some ways of explaining it.

Introduction

Sigurd, Eeg-Olofsson & van de Weijer 2004 investigated the word length distribution of the million-word corpora *Press-65* (Swedish) and *Brown Corpus* (American English), fitting it to the Good distribution. The Good distribution, which is a special case of the so-called Lerch distribution (Johnson, Kemp & Kotz 2005) is described by the formula:

$$f(l) = C \cdot l^a \cdot b^l,$$

where l is the length (in letters), $f(l)$ is the probability of length l, C a normalizing constant, and a and b parameters whose values depend on the particular language.

Fitting the Good distribution to more languages

For the work reported here, the *Regress+* software has been used to fit word type length data from six different European languages to the Good distribution. The data are based on the frequency word lists of the freely available *Leipzig corpora collection*, each sample containing about 100,000 sentences. The languages are English, Finnish, French, Sorbian (a Slavic