Semilinearity is a Syntactic Invariant: A reply to Michaelis and Kracht (1997) *

Rajesh Bhatt
University of Texas at Austin
bhatt@cs.utexas.edu

Aravind Joshi
University of Pennsylvania
joshi@linc.cis.upenn.edu

August 8, 2003

1 Introduction

How fast do sentences of a language grow? If we order the sentences of a language in the order of increasing length then how rapidly can the sentence lengths grow? One simple hypothesis is that the growth is bounded by a constant, i.e., for all $i$, $|S_{i+1}| - |S_i|$ is bounded by a constant, say, $C$, where $|S_k|$ is the length of the sentence $S_k$. The basic intuition is that a sentence of length $l$ can be lengthened by some fixed set of lengths corresponding to the lengths of minimal clauses or phrases, whose lengths are bounded. This is the Constant Growth Property (CGP) that Joshi (1985) first proposed in his paper on the amount of context-sensitivity needed for natural language description.

In that paper Joshi put forward the hypothesis, which we will call the MCSL hypothesis, that natural languages lie in the class of Mildly Context Sensitive Languages (MCSLs). The class of MCSLs is characterized by the following properties: (a) worst case parsing complexity of a language in this class is polynomial,

*We would like to thank the two anonymous reviewers for very helpful comments.
i.e., it is proportional to $n^k$, for some integer $k$ and $n$ is the length of the sentence, (b) grammars for languages in this class can characterize a limited set of patterns of nested and crossed dependencies and their combinations, (c) Context-free languages are properly included in this class, (d) languages in this class have the constant growth property (CGP). The MCSL hypothesis has generated very fruitful research in comparing different linguistic and formal proposals and discovering some equivalences among formalisms, leading to a proper interplay of formal frameworks, substantive linguistic theories, and computational and processing paradigms. Within the context of the MCSL hypothesis, the empirical question of whether the CGP holds of natural languages takes on added significance.¹

The property CGP is a weaker case of the well-known Semilinearity Property (SP) that holds for context-free languages.² SP requires the length of a sentence to be a linear combination of some fixed set of lengths. Before we begin our discussion, we should note that certain formal languages do not possess SP/CGP. e.g. $L_1 = \{a^n | n \geq 1\}$, $L_2 = \{a^{2n} | n \geq 1\}$, $L_3 = \{a^n | n$ is a prime number\} among others. SP/CGP thus severely restricts the class of mildly context-sensitive languages. The cases under discussion in this paper do not discriminate between CGP and SP. We use SP as a cover term to describe the relevant growth properties.

Over the years, since the publication of Joshi (1985) two potential counter examples have been proposed, suggesting that SP may not hold in general for natural languages. These two examples are: (1) Dutch coordination with crossed dependencies (Manaster-Ramer (1987)) and (2) Chinese Number-Names (Radzinski (1991)). Michaelis and Kracht (1997) discuss both these cases, including some of the problems in the argumentation needed to show that SP does not hold for these cases. These problems were already pointed out by Manaster-Ramer.
and Radzinski themselves in their papers. Because of these problems, these two potential counterexamples are not as convincing as they appear at first sight. Michaelis and Kracht therefore propose a new potential counterexample concerning the phenomenon of suffixaufnahme in Old Georgian. They claim that this example provides a stronger argument against SP because it deals with morphosyntactic phenomena and does not depend on the existence of some ‘invisible’ elements, as, for example, in the case of Dutch coordination with crossed dependencies. Our present paper deals with this suffixaufnahme example.

On the basis of data from Old Georgian (Boeder (1995)), Michaelis and Kracht (1997) argue against treating semilinearity as a syntactic invariant. They claim that the phenomena of suffixaufnahme in Old Georgian noun phrases is responsible for making Old Georgian a non-semilinear growth language. We show that (i) Michaelis and Kracht drew an incorrect inference from the data presented in Boeder (1995), and (ii) they ignored certain processes of morphological reduction (haplology). Once these two factors are taken into account, the claim that Old Georgian is a non-semilinear growth language becomes untenable.

2 Michaelis and Kracht’s Claim

Old Georgian permitted possessive NP’s and other arguments of the head noun to appear either pre- or post-nominally. Boeder (1995) argues that the prenominal version is basic and that the postnominal version is derived from the prenominal version.

(1) a. prenominal

[[[[Davit-is] galob-isa] muq{l-ta} ama-t] çartkuma-{{j}}]
David-Gen singing-Gen verse-Pl(Gen) Art-Pl(Pl(Gen)) recitation-Nom
‘the recitation of the verses of the song of David’
b. postnominal, suffixaufnahme

[saidumlo-igasi-supesa-int-m-is-\[\gamma mr-t-isa-\text{-j}-sa-j]]
mystery-Nom Art-Nom kingdom-Gen Art-Gen God-Gen-Gen-Nom
‘the mystery of the kingdom of God’

c. postnominal, suffixaufnahme

[govel-i igisixsi-sat[a-\text{-j}]-m-is]
all-Nom Art-Nom blood-Nom house-Gen-Nom Art-Gen
[Saul-is-\text{-j}]
Saul-Gen-Gen-Nom
‘all the blood of the house of Saul’

When the arguments of the noun appear after it, they display suffixaufnahme: they take up case markings associated with a structurally higher noun. Consider \(\gamma mr-t-isa-j-sa-j\) ‘God-Gen-Gen-Nom’ in (1b). The innermost Gen marking -isa- is what one might expect going from a language like English. The second Gen -jsa- marking comes from the immediately higher noun sasupevel ‘kingdom’ and the highest case marking, the Nominative -j comes from the highest noun saidumlo ‘mystery’.

Similarly, we see in (1c) that the noun at level one of the embedding, saxl ‘house’, has one extra level of case marking, and that the noun at level two of the embedding, Saul, has two extra levels of embedding.

Michaelis and Kracht infer the following pattern from the above data.

(2) General Pattern: \([N_1-\text{-K}_1 \ [N_2-\text{-K}_2-\text{-K}_1 \ [N_3-\text{-K}_3-\text{-K}_2-\text{-K}_1 \ldots \ [N_n-\text{-K}_{n-1-\ldots-\text{-K}_1}]..]]\]

Length: \(n^2 + 3n) / 2\)

a. Level 3:

\(N_1-\text{-K}_1 \ N_2-\text{-K}_2-\text{-K}_1 \ N_3-\text{-K}_3-\text{-K}_2-K_1\)

b. Level 4:

\(N_1-\text{-K}_1 \ N_2-\text{-K}_2-\text{-K}_1 \ N_3-\text{-K}_3-\text{-K}_2-\text{-K}_1 \ N_4-\text{-K}_4-\text{-K}_3-\text{-K}_2-\text{-K}_1\)
If we sum up the number of case-markers in (2), we see that if the embedding is of depth \( n - 1 \) i.e. there are \( n \) nouns, the number of case markers is \( \frac{n(n + 1)}{2} \) and the total number of morphemes in the noun phrase is \( \frac{n(n + 3)}{2} \). Michaelis and Kracht then show that if a language has a productive pattern of the form illustrated in (2), then that language does not have the constant growth property and hence does not have \( \text{SP} \).\(^3\)

We agree with Michaelis and Kracht’s conditional conjecture. However, we show that the antecedent does not hold i.e. Old Georgian does not have a productive pattern of the sort illustrated in (2). Our argument is two-pronged. We first show that the pattern in (2) is not part of the well-formed structures of Old Georgian. Then we show that even if the syntax, in principle, permitted structures of the form in (2), morphological principles would not permit non-constant growth.

3 Boeder’s analysis of Suffixaufnahme

In this section, we show that Michaelis and Kracht erroneously arrive at (2) because they ignored the difference made by Boeder between two distinct varieties of suffixaufnahme: simple suffixaufnahme and multiple suffixaufnahme. Neither variety leads to non-constant growth by itself. It is only an unattested interaction between the two kinds of suffixaufnahme that can lead to non-constant growth.

3.1 Multiple Suffixaufnahme

Multiple Suffixaufnahme is exemplified by (3).
(3) Multiple Suffixaufnahme

a. \( N_1\text{-}K_1 \) \( N_2\text{-}K_2 \) \( N_3\text{-}K_3 \) \( \ldots \) \( N_n\text{-}K_{n-1}\ldots-K_3\text{-}K_2\text{-}K_1 \)

b. [saidumloj \( \text{igi} \) [sasupevel-isa m-is \( \gamma\text{mrt-isa-jsa-j} \)]]
   
   mystery-Nom Art-Nom kingdom-Gen Art-Gen God-Gen-Gen-Nom
   
   ‘the mystery of the kingdom of God’

(3a) shows the general pattern, and (3b) provides an example. It can be shown that if the depth of embedding is \( n - 1 \) i.e. there are \( n \) nouns, then there are \( n + (n - 1) \) case markers. As a result, the total number of morphemes in the noun phrase is \( 3n - 1 \). In other words, multiple suffixaufnahme on its own does not lead to non-constant growth.

Boeder’s particular method of deriving the word order in (3b) is described below. He assumes that the word order in (4) is basic and that the order in (3b) is derived from (4) by a series of rightward movements.

(4) [igi \( \text{[m-is \( \gamma\text{mrt-isa}\)]_{NP_3} \text{sasupevel-isa]}_{NP_2} \text{saidumlo-j]}_{NP_1} \)]
   
   Art-Nom Art-Gen God-Gen kingdom-Gen mystery-Nom
   
   ‘the mystery of the kingdom of God’

The case-stacking arises due to the following facet of Old Georgian case-marking: case-marking appears on the final nominal element in a noun phrase. Typically the final element is the head of the noun phrase. When NP-internal rightward movement takes place, the head of the noun phrase stops being NP-final. While it retains its original case-marking, the new NP-final constituent also receives the original case marking. The steps in the derivation of (3b) from (4) are shown in (5).

(5) a. Basic Structure:

\[
[N_{P_1} \ [N_{P_2} \ [N_{P_3} \ N_3\text{-}K_3 \ N_2\text{-}K_2 \ N_1\text{-}K_1] \ ]]
\]

b. Rightward Movement of \( NP_3 \) inside \( NP_2 \), followed by case assignment:
c. Rightward Movement of NP₂ inside NP₁ followed by case assignment:

\([NP₁ [NP₂ t_{NP₃} N₂-K₂ [NP₃ N₃-K₃-K₂]] N₁-K₁]\)

(\text{the final word order})

### 3.2 Simple Suffixaufnahme

The other kind of suffixaufnahme, which Boeder calls simple suffixaufnahme is less frequent, and moreover is only attested for nominative case. It should be noted that in contrast to genitive case, which is involved in Multiple Suffixaufnahme, nominative case always has a noun phrase-external assigner. So it is not surprising that the two suffixaufnahme processes differ on the basis of case.

(6) exemplifies Simple Suffixaufnahme.

(6) a. Basic Pattern

\(N₁-K₁ \ N₂-K₂-K₁ \ N₃-K₃-K₁ \ldots \ Nₙ-Kₙ-K₁\)

(where \(K₁ = \text{Nom}\))

b. sul-i  sulneleb-isa-j  sakumemvel-ta-j
breath-Nom fragrance-Gen-Nom incense-Pl(Gen)-Nom

‘a whiff of fragrance of incenses’

Like multiple suffixaufnahme, simple suffixaufnahme also does not lead to non-constant growth. The number of morphemes in the NP in (6a) is \(n\) (for the \(n\) nouns) plus \(n\) (for the associated case-markers), and another \(n - 1\) for the extra \(K₁\)’s i.e. a total of \(3n - 1\).

### 3.3 Interactions between the two kinds of Suffixaufnahme

Boeder notes that there are examples which seem to combine both kinds of suffixaufnahme (cf. 7).
Now this case does look like it could involve non-constant growth. Therefore, let us consider on a more general level what it means to put together simple and multiple suffixaufnahme. For ease of exposition, the patterns in both simple and multiple suffixaufnahme are displayed below in (8).

(8) a. Simple Suffixaufnahme:
   Level 3: N₁-K₁ N₂-K₂-K₁ N₃-K₃-K₁
   Level 4: N₁-K₁ N₂-K₂-K₁ N₃-K₃-K₁ N₄-K₄-K₁
   Level 5: N₁-K₁ N₂-K₂-K₁ N₃-K₃-K₁ N₄-K₄-K₁ N₅-K₅-K₁

b. Multiple Suffixaufnahme:
   Level 3: N₁-K₁ N₂-K₂ N₃-K₃-K₂-K₁
   Level 4: N₁-K₁ N₂-K₂ N₃-K₃ N₄-K₄-K₃-K₂-K₁
   Level 4: N₁-K₁ N₂-K₂ N₃-K₃ N₄-K₄ N₅-K₅-K₄-K₃-K₂-K₁

Putting together the two patterns of suffixaufnahme, we get the pattern in (9).

(9) a. Simple and Multiple Suffixaufnahme together:
   Level 3: N₁-K₁ N₂-K₂-K₁ N₃-K₃-K₂-K₁
   Level 4: N₁-K₁ N₂-K₂-K₁ N₃-K₃-K₁ N₄-K₄-K₃-K₂-K₁
   Level 4: N₁-K₁ N₂-K₂-K₁ N₃-K₃-K₁ N₄-K₄-K₁ N₅-K₅-K₄-K₃-K₂-K₁

If we were to look at just Level 3, it might seem that what we have in the general case is the structure in (10).

(10) N₁-K₁ N₂-K₂-K₁ ⋯ Nₙ-Kₙ-Kₙ₋₁⋯-K₂-K₁
This structure, as Michaelis and Kracht have pointed out, leads to non-constant growth. However, once we look at structures which have further embedding, it becomes clear that the generalization assumed by Michaelis and Kracht was not the correct one. Instead the correct generalization seems to be the one in (11).

(11) \[ N_1-K_1 \ N_2-K_2-K_1 \ \ldots N_i-K_i-K_1 \ \ldots N_n-K_n-K_{n-1} \ldots-K_2-K_1 \]

The NP with \( n \) levels of embedding in (11) has \( 4n - 3 \) morphemes. Clearly this structure does not lead to non-constant growth.\(^5\)

(12) General pattern of co-occurring Simple and Multiple Suffixaufnahme:

\[ N_1-K_1 \ N_2-K_2-K_1 \ \ldots N_i-K_i-K_1 \ \ldots N_n-K_n-K_{n-1} \ldots-K_2-K_1 \]

length: \( 4n - 3 \)

a. Simple and Multiple Suffixaufnahme together:

Level 3: \( N_1-K_1 \ N_2-K_2-K_1 \ N_3-K_3-K_2-K_1 \)

Level 4: \( N_1-K_1 \ N_2-K_2-K_1 \ N_3-K_3-K_1 \ N_4-K_4-K_3-K_2-K_1 \)

Level 5: \( N_1-K_1 \ N_2-K_2-K_1 \ N_3-K_3-K_1 \ N_4-K_4-K_1 \ N_5-K_5-K_4-K_3-K_2-K_1 \)

3.3.1 The absence of recursive interactions

The reader might wonder about the possibility of repeated applications of simple suffixaufnahme i.e. in addition to simple suffixaufnahme just applying at the top level, why does simple suffixaufnahme not apply recursively, at each level. Then we would indeed end up with Michaelis and Kracht’s (10).

We do not know the ultimate answer to this question - but we do have a partial answer based on an observation of Boeder’s. Boeder points out that simple suffixaufnahme is limited to Nominative case. Nominative case is always assigned from outside the Noun Phrase. All the other potential candidates for recursive simple suffixaufnahme involve cases other than Nominative. Conse-
quently, simple suffixaufnahme is limited to the top-most level which is the only level at which Nominative may be assigned.

4 The Role of Haplology

We have shown so far that the phenomena of suffixaufnahme in Old Georgian does not lead to non-constant growth. Closer examination reveals the growth in terms of length to be even more limited. This is because of the phenomena of Haplology and other morphological sequence constraints (Dench and Evans 1988:35-43).

The term haplology is used to describe processes by which a whole syllable is deleted before or after a phonetically similar or identical syllable. The haplological process may be obligatory (13a) or optional (13b, c).

(13) (from Hock 1986:109)

a. Latin: *nutritrix > nutritrix ‘nurse’

b. Homeric Greek: amphiphoreus > amphißoreús ‘two-handled pitcher’

c. Latin: trierarchus > trierarchus/triararhchus ‘captain of a triera’

Boeder notes that haplology is obligatory in the case of the repeated oblique plural genitive marker -ta and optional in the case the repeated singular genitive marker -isa.

(14) a. Genitive Haplology

i. z-isa kac-isa-jsa
   son-Gen man-Gen-Gen
   ‘the son of man’

ii. z-isa kac-isa
    son-Gen man-Gen
    ‘the son of man’
b. Plural Oblique Haplology

i. *kar-ta kalak-ta-ta
door-Pl(Obl) city-Pl(Gen)-Pl(Pl)

ii. kar-ta kalak-ta
door-Pl(Obl) city-Pl(Gen)
‘the gates of the cities’

In addition, there seem to be no instances of three consecutive repetitions of the same case marker i.e. there are no instances of Gen-Gen-Gen in the examples cited by Boeder. Our speculation is that such kind of case-stacking is also ruled out by a haplological constraint. If these speculations are in the right direction, the pattern of growth in (12) reduces by haplology to (15).

(15) General pattern of co-occuring Simple and Multiple Suffixaufnahme:

\[ N_1-K_1 \ N_2-K_2-K_1 \ \ldots \ N_i-K_i-K_{i-1} \ \ldots \ N_n-K_n-K_{n-1} \ldots -K_2-K_1 \]

length: \(4n - 3\)

Assume \(K_1 = \text{Nom}\), and \(K_2, \ldots, K_n = \text{Gen}\), and haplological constraints rule out *Gen-Gen-Gen

Reduced pattern: \(N_1-K_1 \ N_2-K_2-K_1 \ \ldots \ N_i-K_i-K_{i-1} \ \ldots \ N_n-K_n-K_{n-1}-K_1 \)

length: \(3n\)

a. Simple and Multiple Suffixaufnahme together:

Level 3: \(N_1-K_1 \ N_2-K_2-K_1 \ N_3-K_3-K_2-K_1\)

Level 4: \(N_1-K_1 \ N_2-K_2-K_1 \ N_3-K_3-K_1 \ N_4-K_4-K_3-\phi-K_1\)

Level 5: \(N_1-K_1 \ N_2-K_2-K_1 \ N_3-K_3-K_1 \ N_4-K_4-K_1 \ N_5-K_5-K_4-\phi-\phi-K_1\)

In fact, we expect that even if Old Georgian in principle permitted non-constant growth in the pattern assumed by Michaelis and Kracht, in reality due to the haplological constraints the growth would still be constant. The pattern of growth assumed by Michaelis and Kracht is shown in (2) repeated here as (16).
(16) General Pattern: \([N_1-K_1 [N_2-K_2-K_1 [N_3-K_3-K_2-K_1 \ldots [N_n-K_n-\ldots-K_1]..]]]\)

Length: \((n^2 + 3n)/2\)

a. Level 3:
\(N_1-K_1 N_2-K_2-K_1 N_3-K_3-K_2-K_1\)

b. Level 4:
\(N_1-K_1 N_2-K_2-K_1 N_3-K_3-K_2-K_1 N_4-K_4-K_3-K_2-K_1\)

c. Level 5:
\(N_1-K_1 N_2-K_2-K_1 N_3-K_3-K_2-K_1 N_4-K_4-K_3-K_2-K_1 N_5-K_5-K_4-K_3-K_2-K_1\)

We predict that this pattern, if it existed, would reduced to (17).

(17) Reduced Pattern: \([N_1-K_1 [N_2-K_2-K_1 [N_3-K_3-K_2-K_1 \ldots N_i-K_i-K_{i-1}-K_1 \ldots [N_n-K_n-K_{n-1}-K_1]..]]\)

assuming \(K_1 = \text{Nom}\), and \(K_2, \ldots, K_n = \text{Gen}\)

Length: \(4n - 3\)

a. Level 3:
\(N_1-K_1 N_2-K_2-K_1 N_3-K_3-K_2-K_1\)

b. Level 4:
\(N_1-K_1 N_2-K_2-K_1 N_3-K_3-K_2-K_1 N_4-K_4-K_3-K_1\)

c. Level 5:
\(N_1-K_1 N_2-K_2-K_1 N_3-K_3-K_2-K_1 N_4-K_4-K_3-K_1 N_5-K_5-K_4-K_1\)

This reduced pattern does not involve non-constant growth.

5 Summary

We have shown that contrary to Michaelis and Kracht’s claim the syntax of Old Georgian did not permit non-constant growth. Michaelis and Kracht were led
to their conclusion due to an incorrect assumption about the nature of suffix-aufnahme in Old Georgian.\textsuperscript{6} We have also shown that there are morphological processes in language that militate against arbitrary stacking of case morphology. These processes lead to constant growth even if they are presented with structures that by themselves would lead to non-constant growth.

References


1 A reviewer has commented that CGP is interestingly different from other proposed linguistic universals in that it is stated in terms of the number of countable morphemes/words in a sentence and not in terms of structural categories (NP, VP, etc.). Despite this, the reviewer notes, CGP is connected with structural properties in a deep way. It distinguishes between the classes of formalisms/languages that seem right for natural languages (MCSLs) and more powerful formalisms/languages.

2 CGP holds for the languages generated by Tree-Adjoining Grammars (TAG), certain versions of Combinatory Categorial Grammars (CCG), Head grammars (HG) (Joshi, Vijay-Shanker and Weir 1991), Stabler’s version of minimalist grammars (Stabler 1997), and the TAG formulation of minimalist grammars according to Frank (Frank 2002). SP has been shown to hold for the first three (see Joshi et al. (1991) for TAG, CCG, and HG). The status for the last two is not known but we believe that SP also holds for them too.

3 More specifically they show that a language obtained by intersecting Old Georgian with a specific regular set does not have SP. This is shown by the fact this
intersection language grows in a non-constant growth manner.

4 The derivation shown in (5) is a bottom-up derivation. Boeder offers a top-down derivation. The differences between the two are independent of the discussion here.

5 For levels of embedding greater than three, the results of Boeder’s derivation of examples like (7) diverges from the results proposed by us in this section. Some word orders generated by Boeder’s derivation are shown below.

i. The General Case:
   \[ N_1 \cdot K_1 \cdot N_2 \cdot K_2 \cdot K_1 \ldots N_i \cdot K_i \cdot K_{i-1} \cdot K_1 \ldots N_n \cdot K_n \cdot K_{n-1} \cdot K_1 \]
   length: \( 4n - 3 \)
   a. Level 3:
      \[ N_1 \cdot K_1 \cdot N_2 \cdot K_2 \cdot K_1 \cdot N_3 \cdot K_3 \cdot K_2 \cdot K_1 \]
   b. Level 4:
      \[ N_1 \cdot K_1 \cdot N_2 \cdot K_2 \cdot K_1 \cdot N_3 \cdot K_3 \cdot K_2 \cdot K_1 \cdot N_4 \cdot K_4 \cdot K_3 \cdot K_1 \]
   c. Level 5:
      \[ N_1 \cdot K_1 \cdot N_2 \cdot K_2 \cdot K_1 \cdot N_3 \cdot K_3 \cdot K_2 \cdot K_1 \cdot N_4 \cdot K_4 \cdot K_3 \cdot K_1 \cdot N_5 \cdot K_5 \cdot K_4 \cdot K_1 \]

We speculate that these orders are not the desired orders. While they do display simple suffixaufnahme (note the repeated \( K_1 \)), they do not display multiple suffixaufnahme on the final noun (note the missing \( K_3 \) and \( K_2 \)). Further, there is an additional unexpected suffixing of the case assigned by the two-levels higher noun. Unfortunately, there is no data to confirm or disconfirm our speculation in the corpus of examples in Boeder.

The point we would like to stress is that (i) Boeder’s system does not lead to non-constant growth, and (ii) the expected pattern, shown in (9) and repeated
here as (12), also does not lead to non-constant growth. In fact the number of morphemes at level $n$ of embedding is the same ($= 4n-3$) irrespective of which of (i) or (12) is actually attested.

A reviewer suggests that the facts concerning the distribution of case-marking over each element of a conjunction discussed in Boeder (1995):160-161 could be used to construct a new argument that Old Georgian is not semi-linear. We examined the data in question and did not see how they would support an argument against CGP (and therefore against SP)—in fact it is clearly seen that they do not involve non-constant growth.