Glottometrics 12

2006

RAM - Verlag

Glottometrics

Glottometrics ist eine unregelmäßig erscheinende Zeitschrift für die quantitative Erforschung von Sprache und Text
Beiträge in Deutsch oder Englisch sollten an einen der Herausgeber in einem gängigen Textverarbeitungssystem (vorrangig WORD) geschickt werden
Glottometrics kann aus dem **Internet** heruntergeladen, auf **CD-ROM** (in PDF Format) oder in **Buchform** bestellt werden

Glottometrics is a scientific journal for the quantitative research on language and text published at irregular intervals
Contributions in English or German written with a common text processing system (preferably WORD) should be sent to one of the editors
Glottometrics can be downloaded from the **Internet**, obtained on **CD-ROM** (in PDF) or in form of **printed copies**

Herausgeber – Editors

Bestellungen der CD-ROM oder der gedruckten Form sind zu richten an
Orders for CD-ROM´s or printed copies to

RAM-Verlag RAM-Verlag@t-online.de

Herunterladen / Downloading: http://www.ram-verlag.de

Die Deutsche Bibliothek – CIP-Einheitsaufnahme

Glottometrics. –12 (2006) –. – Lüdenscheid: RAM-Verl., 2006
Erscheint unregelmäßig. – Auch im Internet als elektronische Ressource unter der Adresse http://www.ram-verlag.de verfügbar.-
Bibliographische Deskription nach 12 (2006)
ISSN 1617-8351

Glottometrics 12, 2006, 11-15

Script complexity revisited

Carsten Peust, Konstanz[1]

Abstract: A simple method for quantifying the complexity of graphical signs is suggested. The complexity is defined as the number of crossing points which can maximally be achieved with an overlapping straight line. In signs composed of several disconnected elements, the complexity is to be computed for each element seperately. This method is compared with another complexity measure serving a similar purpose recently proposed by Altmann.

It is intuitively clear that graphical signs can have different degrees of *complexity*: Q is obviously more complex than I, and A is more complex than A. Altmann made a proposal for numerically measuring this intuitive category recently in this journal (Altmann 2004). His proposal (in the following *composition method*) is to split a sign into basic elements, for which he defines specific costs (point: costs 1; straight line: costs 2; arch: costs 3). In addition, three types of contacts are to be considered, again involving costs (continuous: costs 1; crisp: costs 2; crossing: costs 3). The complexity of a sign is the sum of the costs of all of its basic elements as well as contacts. Altmann finds that the complexity counts achieved by his system agree well with our intuitive understanding of sign complexity.

Altmann's composition method indeed turns out to be a useful way of quantifiying sign complexity, but a slight feeling of dissatisfaction remains. Two quite heterogeneous principles are at work (elements and contacts), together with six cost values which have to be defined arbitrarily. In this paper, an alternative method is proposed which makes use only of a single principle and therefore involves fewer arbitrary assumptions, is applicable even more rapidly and unambiguously, and still reflects well our intuitive idea of sign complexity (in the following *intersection method*). I wish to thank Gabriel Altmann who has first drawn my attention to the topic and who generously offered me the opportunity to publish a counterproposal in this journal.

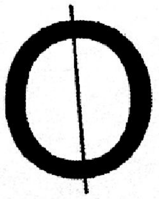
Assume that a straight line overlaps with the sign to be measured. This results in crossing points, whose number will vary for different positions and rotations of the line. What we will consider is the number of crossing points which can be maximally achieved with an optimal placement of the line over the graphical sign. I thus propose the following definition:

Rule 1: *The complexity of a sign is the maximal number of crossing points that can be achieved with a straight line.*

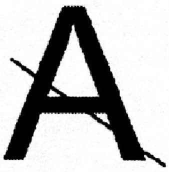
Instead of counting crossing points, it is an equivalent possibility to count the number of black/white-transitions and divide it by two. A few examples will suffice to illustrate the idea.

O has a complexity of 2 because there are, at maximum, two crossings or four black/white-transitions:

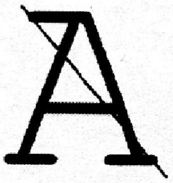
[1] Address correspondence to: Carsten Peust, Bücklestr. 68a, D-78467 Konstanz, Germany. E-mail: cpeust@gmx.de



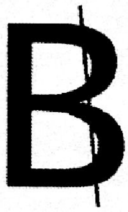
A has a complexity of 3:



A has a complexity of 5:



In applying this method, the sign should be considered as being composed of lines with minimal width. The sign **B** arrives at a complexity of 4 because, assuming it as a shape of minimal thickness, a straight line can be placed so as to cross each of its two arches twice, which results in four crossings or eight black/white-transitions:



It seems desirable to achieve the same complexity also for the same sign in a bolder type (e.g. **B**), although eight black/white-transitions are not actually possible here because the outline is too thick.

The following table illustrates which complexity values are reached for the majuscles of the Latin alphabet in the fonts Arial and Courier New according to both algorithms:

Arial	*Composition Method*	*Intersection Method*	**Courier New**	*Composition method*	*Intersection Method*
A	12	3	A	22	5
B	16	4	B	16	4
C	7	2	C	11	3
D	9	2	D	9	3
E	14	3	E	26	5

F	10	3	F	22	5
G	15	3	G	15	4
H	10	3	H	26	5
I	2	1	I	10	3
J	3	2	J	7	3
K	10	3	K	26	6
L	6	2	L	14	3
M	14	4	M	26	6
N	10	3	N	20	5
O	8	2	O	8	2
P	9	3	P	13	4
Q	13	3	Q	21	4
R	14	4	R	22	6
S	15	3	S	23	5
T	6	2	T	18	3
U	3	2	U	11	3
V	6	2	V	14	3
W	14	4	W	22	5
X	7	2	X	23	4
Y	8	2	Y	20	4
Z	10	3	Z	18	3

It appears that a higher complexity in one system normally implies a higher complexity in the other: Both methods are roughly equivalent. Based on the given set of 52 test signs, the following correspondency table can be set up:

Intersection Method	1	2	3	4	5	6
Composition Method	2	3-9	7-18	13-23	20-26	22-26

On the other hand, there do exist sign couples for which the two methods lead to contradictory results. Some extreme cases are shown in the next table. They are arranged so that the left-hand sign is less complex and the right-hand sign more complex according to the composition algorithm, but vice versa according to the intersection algorithm:

Sign couple		Composition complexity		Intersection complexity	
J	D	7	9	3	2
P	T	13	18	4	3
N	X	20	23	5	4
R	H	22	26	6	5

The crucial question is whether the signs in the left-hand column are the "simpler" ones (which would favour the composition method) or the more "complex" ones (which would favour the intersection method). A spontaneous judgement may seem difficult, but it is my perception that the advantage lies slightly on the side of the intersection method.

Another difference is that Altmann's composition method leads to a much finer gradation of complexity than does the intersection method proposed here. This may be desirable. On the other hand, I believe that these gradations are on the whole not confirmed by our intuitive notions of sign complexity. Let us inspect all those signs which, with the intersection algorithm, are uniformly assigned the complexity of 3, sorted in increasing (1) and decreasing (2) order of their composition complexity, which varies as widely as from 7 to 18:

(1): J P D F H K N Z I C U A Q E L V G S T Z
(2): Z T S G V L E Q A U C I Z N K H F D P J

Can it be recognized that the complexity increases in (1) and decreases in (2), which is what Altmann would postulate? I do not think so. Let us examine two other sequences of signs, namely signs for which both methods unanimously predict an increase (3) or a decrease (4) of complexity:

(3): I U H B F M
(4): M F B H U I

There is no doubt in this case as to which row has increasing and which one has decreasing complexity.

There is, however, a considerable systematic advantage of Altmann's composition method compared with the intersection method as presented up to now. Altmann's system allows for the simple addition rule *C(a)* + *C(b)* = *C(a+b)*: The complexity of a sign group is the sum of the complexities of its individual elements. The composition complexity can thus be plausibly applied also to greater entities such as words or even texts. With the intersection method, however, the value for a word would normally fall drastically below the sum of its component values because most letters would not be crossed optimally or not be touched at all by a single line.

This theoretical problem leads to unintuitive results even in the computation of single letters as soon as they become more complex. Consider the Arabic letters س (*s*) and ش (*š*). س has an intersection complexity of 6 (cut twice each through the bottom of all three arches). Although the shape of ش is clearly derived from س by the addition of diacritical points, these points do not affect its intersection complexity, as defined up to now, because a straight line cannot be placed optimally so as to include any of them. It is clearly counterintuitive that both signs should have the same complexity.

As another example, consider the Korean group 음 (*ym*) (the readers are asked to take the bottom element as a simple rectangle with no serifes for the sake of this argument). The complexity will be 5 (cut through all three elements in vertical direction). The similar group 임 (*im*), which differs only in the arrangement of its elements and is likely to be regarded as identical in complexity by many, achieves, however, a complexity of only 4 (again with a vertical line).

In order to remedy such unintuitive results, I wish to posit the following additional rule:

<u>Rule 2</u>: *The complexity of a graphical cluster consisting of several disconnected components must not be computed with a single straight line. Instead, its complexity is defined as the sum of the complexities of its components.*

This ensures that the attractive additive behaviour of Altmann's composition complexity can be retained also in our system. The complexity of ش, a sign composed of four components separated by white space, will now be 6+1+1+1 = 9; the complexity of both 음 and 임 will be 2+1+2 = 5. Our additional rule does not change any of the values of the Latin capitals as discussed above since none of them is composed of disconnected elements.

What range of values will be achieved for sign complexity by applying the intersection method proposed here? As we saw, the values fall between 1 and 6 for Latin majuscles; the results are similar for the minuscles. In Arabic, values are largely the same, with an exceptionally high value of 9 for the already discussed letter ش (*š*) with three diacritics. In the Tamil alphabet we reach complexities up to 7: இ (*i*) and 8: ண (*ṇ*) (I am neglecting here the composite syllabic groups which can be even more complex). In Chinese, as might be expected, these values are outnumbered even by very familiar signs such as 瞧 (*qiáo* "to look", complexity 16), 餐 (*cān* "food", complexity 17), 罐 (*guàn* "tin, box", complexity 19), or 露 (*lù* "dew", complexity 20).

Reference

Altmann, G. (2004). Script complexity. *Glottometrics 8, 68-74.*