Principles Aiding in Reading Abbreviations in OldGeorgian and Latin

Hoenen, Armin
hoenen@em.uni-frankfurt.de
Goethe Universität Frankfurt, Germany

Samushia, Lela
samushia@em.uni-frankfurt.de
Goethe Universität Frankfurt, Germany

Introduction

A project on Georgian Epigraphy [1] in the context of which this article is grounded brought our attention to two phenomena which seem inseparably intertwined with the epigraphic record: abbreviations and gaps. In Old Georgian, abbreviations are very prominent and yet very heterogeneous. That is more so than in other (historical) languages. One finds many abbreviations for the same word, compare Boeder (1987). In this article, we assess the research question why this is so by trying to highlight properties of Old Georgian abbreviations, which can also be used as an aid to read them. We use Old Georgian abbreviations in inscriptions and manuscripts from the Titus web-site, Gippert (1995) under more entered within the Georgian National Corpus [2]. In order to take a somewhat wider perspective and enable data hungry technologies to aid in the analyses of the aforementioned phenomena [3], we decided to compare the Georgian record with one of the largest digital epigraphic records available, classical Latin, for which we use data from the Epigraphic database Heidelberg [4]: last accessed December 2016. The Latin data is several orders of magnitude larger than the Georgian one. While Latin contains roughly 70,000 inscriptions, Georgian features 91, so we decided to additionally analyse abbreviations in Old Georgian manuscripts. We ended up with roughly 4,000 abbreviations for Old Georgian (roughly 1,100 from the inscriptions) and roughly 170,000 for Latin.

There are several partly overlapping typologies of abbreviations, compare Marchand (1969); Kreidler (1979); McArthur (1988); McArthur and McArthur (1992); Rúa (2004); Driscoll (2009). (Carroll, 2003, p.205) remarks: ”Unfortunately, universally accepted standards for many abbreviations and acronyms do not exist”.

Position

The first investigation concerns the position of the letters in the extension, which are maintained in the abbreviation. That is <precip> as an abbreviation for <precipitation>, will be converted into 1-2-3-4-5-6. We look at the type and token levels for Old Georgian and Latin, see Table 1.

Text Type	p_0=first	p_l=last	Suspension	Contraction
Old Georgian Inscriptions	0.998	0.895	0.037	0.158
Old Georgian Manuscripts	1.0	0.814	0.185	0.344
Old Georgian Inscriptions (types)	0.998	0.902	0.012	0.076
Old Georgian Manuscripts (types)	1.0	0.979	0.016	0.166
Latin Inscriptions	0.998	0.037	0.424	0.003
Latin Inscriptions (types)	0.987	0.183	0.12	0.01

Table 1: Proportions of abbreviations starting in the first letter (first column), ending in the last (second column), being suspensions (third column) or contractions (fourth column). 'types' here counts each abbreviation - expansion tuple uniquely.

We find that abbreviations usually start in the first letter although this might be part of a prefix in both Latin and Old Georgian. This allows for keeping parafoveal preview information intact, see for instance Slattery et al. (2011); Rayner et al. (2012). Chanceaux et al. (2013) summarizing Dandurand et al. (2011) find that ”initial letters provide more information with respect to word identity than any other letter position”. Initital letters are together with the last letter the ”most visible” letters of a word, Dandurand et al. (2011), which means under more that due to the adjacent spaces, they can more easily be recognized.

Secondly, Old Georgian uses more contractions (abbreviations by first and last letter, a Christian abbreviation tradition which entails the question of how to spiritually correctly contract affixed words inducing a dilemma once many affixes are present), Latin suspensions (abbreviation by the first n letters. Figure 1 shows the patterns of occurrence of positions, given a certain word length. The y-axis represents the percentage of abbreviations (regardless of their lengths) that contain the position specified on the x-axis. In a normalized plot, all wordlengths from 4 to 11 are plotted together. For

Old Georgian there is a gap after some first letters. We conjecture that the end of the word stem is most unlikely to occur in abbreviations. While in Latin, suffixes are often left out, in Georgian, which has an agglutinative morphology, this would lead to considerable difficulties in relating the actual word to the context since some suffixes carry information which in Latin are expressed by independent pre- or postpositions.

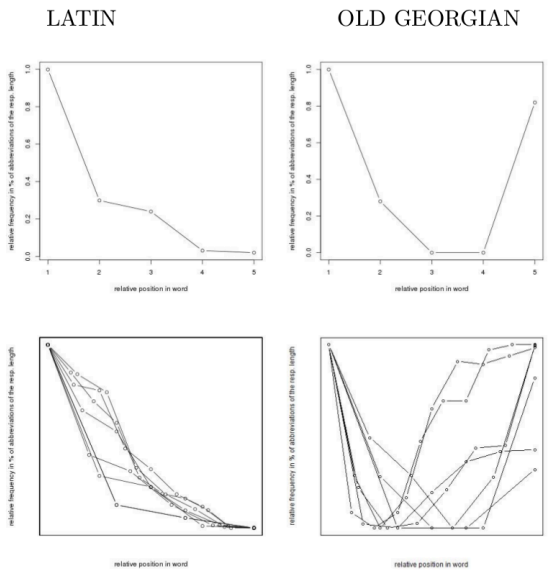

Figure 1: Letter position incidence in abbreviations. First row: for words of length 5 (24, 569 data points in Latin and 401 in Old Georgian). Second row: for words of length 4 to 11 (overlay). In the upper left graphic, for all abbreviated words of length 5 in Latin, the first letter was always present in the abbreviation, the second in roughly 30% of the cases, the third in slightly less cases, the fourth and fifth almost never. In Old Georgian, an alternation with 2 dominant patterns appears: one for shorter words with probably one suffix and one for words with 2 suffixes which then have more letters towards the relative middle.

Vowels & Consonants

Now, we look at how many vowels and consonants are retained in abbreviations, see Table 2. As one can see, in all contexts the ratio of vowels in the abbreviations is clearly lower than in the extensions. Information theory, see for instance Shannon (1948) provides a stable basis for the interpretation of these results. a) the class of vowels is smaller than the class of consonants and b) vowels are much more frequent not only because of a) but also because they occur in syllable nuclei and consequently a single vowel, yet not a single consonant can constitute a syllable. [5] [6] Thus, vowels are less informative and hence more easily guessable than consonants. Finally, we analysed our larger Latin data in more detail. We extract the frequency based rankings of letters and test their correlation with the ranking through probability of being retained in an abbreviation. The Spearman rank correlation suggests, that there is a significant correlation between the frequency of a consonant and its probability to be retained in an abbreviation (p-value = 0.002655, # 0.622807). For vowels p is above 0.01, and # is negative at -0.2 suggesting that for vowels, which are all quite frequent, there is no clear effect. Carreiras et al. (2009) find that vowels do not yield priming effects for recognition in opposition to consonants, which would nicely align with this finding. Thus, an inf requent consonant grapheme is more probably retained. One may deduct that consonants that are left out - if any - are more frequent than the ones retained. Knowing this and knowing that the first portion of the word root is probably represented in the abbreviation decreases the number of possible interpretations ideally by this token helping to decipher abbreviations.

Text Type	**Proportion** of vowels in extension	**Proportion** of vowels in abbreviation
Old Georgian Inscriptions	0.408	0.276
Old Georgian Manuscripts	0.42	0.275
Old Georgian Inscriptions (types)	0.41	0.271
Old Georgian Manuscripts (types)	0.412	0.347
Latin Inscriptions	0.466	0.337
Latin Inscriptions (types)	0.446	0.392

Table 2: Proportions of vowels in extensions and abbreviations. Note, that the inverse (1-value) is the proportion of consonants. Here, vowels and consonants are measured as vowel and consonant characters, which for both Latin and Old Georgian largely coincides with their phonemic class due to both writing systems being rather shallow.

Possibilities for the Abbreviator

Reading abbreviations may be understood even better when taking their generation into account. One hypothesis assumes, that someone, who wants to abbreviate a word, holds in mind all possible abbreviations and chooses from them. We try to validate this, counting all possible abbreviations per word length given suspension and contraction. This can also help estimate algorithmic complexity in abbreviating.

As we have empirically observed but not explicitly stated so far and implicitly assumed, a condition for a valid abbreviation may be that the indices of the letters must

maintain ascending order, or, in other words, the original sequence of the letters is not permuted. If so, for each word length *w* and abbreviation length *a* there are { *w* \ *choose a* } [7] possible abbreviations, since for each distinct combination only one can maintain the ascending order of elements. Now in order for understanding how many possible abbreviations there are for each word we need to add up values for all different *a*, where $a < w$. [8]

By simply using this binomial coefficient one can compute the numbers of possible combinations (for numbers until 10, see Table 3 leftmost numbers in cells) which gives the inner portion of Pascal's triangle. However, the outer 1s are missing, since the extension itself is no abbreviation and neither is a sequence of zero letters, the sum is thus 2^w - 2. The increase in possibilities is linear and quick, for a word of length 15, there are 32,766 possibilities how it could be abbreviated.

If we additionally fix the first letter, we count only all combinations containing element '1', which must then be{ *w-1* \choose *a-1* }, since we have fixed the first letter and from the remaining *w - 1* letters, we can choose any *a - 1* remaining elements of the abbreviation. This restricts the possible numbers already considerably. Overall, we halve possibilities, so 2^{w-1} - 1 becomes the sum formula, which would still leave us with 16,383 possibilities for a word of 15 letters.

In case of a contraction, one fixes the first and the last letter. Then again, results are halved with one letter long abbreviations excluded, hence the sum relates to *w* by 2^{w-2} - 1. For a word of length 15 still 8,191 possibilities would be left.

Numbers remain so high towards the end of the scale that it seems improbable that someone who abbreviates a word be aware of all of the possibilities simulateneously at decision time. One also sees that contraction is quite effective in restricting possible abbreviations and might thus considerably speed-up abbreviating and decipherment/ reading of abbreviations.

	1	2	3	4	5
2	2/1/0				
3	3/1/0	3/2/1			
4	4/1/0	6/3/1	4/3/2		
5	5/1/0	10/4/1	10/6/3	5/4/3	
6	6/1/0	15/5/1	20/10/4	15/10/6	6/5/4
7	7/1/0	21/6/1	35/15/5	35/20/10	21/15/10
8	8/1/0	28/7/1	56/21/6	70/35/15	56/35/20
9	9/1/0	36/8/1	84/28/7	126/56/ 21	126/70/ 35
10	10/1/0	45/9/1	120/36/8	210/84/ 28	252/126/ 56

	6	7	8	9	#
2					2/1/0
3					6/3/1
4					14/7/3
5					30/15/7
6					62/31/15
7	7/6/5				126/63/ 31
8	28/21/15	8/7/6			254/127/ 63
9	84/56/35	36/28/21	9/8/7		510/255/ 127
10	210/126/ 70	120/84/ 56	45/36/28	10/9/8	1022/ 511/255

Table 3: Numbers of possible abbreviations per word length / when first letter is fixed / when first and last letters are fixed. Rows have word length, columns abbreviation length.

Discussion and Conclusion

Various analyses (not only the presented) conducted have shown two properties of abbreviations which can help understand and read Old Georgian and Latin abbreviations.

1. the overwhelming majority of abbreviations contains the first letter of the extension
2. morphological type is an important factor for abbreviating, for Old Georgian the end of the word stem is likely to disappear whereas suffixes tend to be represented, the last letter being likely for the contraction principle
3. for Latin we found that consonants are more likely to occur in an abbreviation than vowels, less frequent consonants more than frequent ones

Combinatorial evidence suggests that keeping in mind all possible abbreviations even in case of contraction is unlikely for longer words. Considering that Old Georgian as an agglutinative language produces long words this may partly explain why there is larger variety in the abbreviation landscape, since in each abbreviation process another possible abbreviation may have surfaced.

Fußnoten

1. https://www.cedifor.de/en/cedifor/current-pilot-projects/pilot-projects/digitale-erschliessung-epigraphischer-denkmaeler
2. http://titus.fkidg1.uni-frankfurt.de/texte/etcg/cauc/ageo/inscr/carcera/carce.htm/ and http://gnc.gov.ge/gnc/page
3. See also Hoenen & Samushia (2016)
4. http://edh-www.adw.uni-heidelberg.de/home

5. In principle, all but sign languages should make use of vowels, at least empirically in the Word atlas of language structures, Dryer and Haspelmath (2013) (see wals.info), there are no languages with less than 2 vowel qualities. Furthermore the lowest consonant vowel ratio in 563 languages has still more consonants (10) than vowels (9), see Maddieson (2013b,a).
6. A third observation on vowels could be related: vowels are synchronically (dialects) and diachronically (language change) less stable.
7. We write formulas in LaTeX syntax in order to avoid for simple formulae to be given too much space.
8. Note, that an abbreviation can be understood as a k-skip-n-gram, Guthrie et al. (2006). The first numbers in the summing column of the table then coincide with the numbers (sum of) of all possible k-skip-n-grams for w (and all possible n and k).

Bibliography

Boeder, W. (1987). Versuch einer sprachwissenschaftlichen Interpretation der altgeorgischen Abkürzungen. In: *Revue des études géorgiennes et caucasiennes* 3:33–81.

Carreiras, M., Duñabeitia, J. A., and Molinaro, N. (2009). Consonants and vowels contribute differently to visual word recognition: Erps of relative position priming. I: *C erebral Cortex* 19(11):2659–2670.

Carroll, J. (2003). *Oxford handbook of computational linguistics*. Oxford University Press.

Chanceaux, M., Mathôt, S., and Grainger, J. (2013). Flank to the left, flank to the right: Testing the modified receptive field hypothesis of letter- specific crowding. In: *Journal of Cognitive Psychology* 25(6):774–780.

Dandurand, F., Grainger, J., Duñabeitia, J. A., and Granier, J.-P. (2011). On coding non-contiguous letter combinations. In: *Frontiers in psychology* 2:136.

Driscoll, M. (2009). Marking up abbreviations in old norse-icelandic manuscripts. In: *Medieval Texts–Contemporary Media*. Ibis.

Dryer, M. S. and Haspelmath, M., (eds) (2013). WALS Online. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Gippert, J. (1995). TITUS. Das Projekt eines indogermanistischen Thesaurus ("TITUS. The project of an Indo-European thesaurus"). In: *LDV- Forum* 12(2):35–47.

Guthrie, D., Allison, B., Liu, W., Guthrie, L., and Wilks, Y. (2006). A closer look at skip-gram modelling. In: *Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC-2006):* 1–4.

Hoenen, A. and Samushia, L. (2016). Gepi: An Epigraphic Corpus for Old Georgian and a Tool Sketch for Aiding Reconstruction, In: *JLCL* 31 (2):25–38.

Kreidler, C. W. (1979). Creating new words by shortening. In: *Journal of English Linguistics* 13(1):24–36.

Maddieson, I. (2013a). WALS, Consonant-Vowel Ratio. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Maddieson, I. (2013b). WALS, Vowel Quality Inventories. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Marchand, H. (1969). The categories and types of present-day English word formation: A synchronic-diachronic approach. Beck.

McArthur, T. (1988). The cult of abbreviation. In: *English Today* 4(03):36–42.

McArthur, T. B. and McArthur, F. (1992). The Oxford companion to the English language. Oxford University Press.

Rayner, K., Pollatsek, A., Ashby, J., and jr. Clifton, C. (2012). Psychology of Reading. Psychology Press, New York/Hove.

Rúa, P. L. (2004). Acronyms & co.: A typology of typologies= acrónimos y cía: una tipología de tipologías. In: *Estudios Ingleses de la Universidad Complutense* 12:109–129.

Shannon, C. E. (1948). A mathematical theory of communication. In: *Bell System Technical Journal* 27:379–423.

Slattery, T. J., Schotter, E. R., Berry, R. W., and Rayner, K. (2011). Parafoveal and foveal processing of abbreviations during eye fixations in reading: making a case for case. In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 37(4):1022.