Estimating the Distinctiveness of Graphemes and Allographs in Palaeographic Classification

Noga Levy, Lior Wolf, Nachum Dershowitz, Peter Stokes

Introduction

Within the discipline of palaeography, the ‘morphological’ approach tries to describe the letter-shape as a whole, so a letter may be described as a ‘Caroline *a*’ or as an ‘insular *r*’. Aspects of this approach are visible in almost all palaeographical handbooks, particularly those that provide alphabets or selections of letter-forms. An example is Albert Derolez’s, *Palaeography of Gothic Manuscript Books,* which also provides a useful discussion of morphology as a palaeographical method.

Commonly, a morphological system of descriptors contains two main categories: One category is the grapheme or perhaps, more correctly, character: the letter as an abstract entity but with physical form, such as *a, æ*, or a single punctuation mark. The second category is the allograph, namely, a particular way of writing the letter; typical examples include ‘Caroline’ or ‘insular’.

A key question, given any system of descriptors, is to evaluate the relative importance of each of the components of this system. For example, it is well known among palaeographers that the grapheme *a* is very distinctive for late Anglo-Saxon minuscule (Ker, 1957; Dumville, 1988; Stokes, 2005); however, subjective evaluation of distinctiveness could potentially be misleading. It is therefore quite useful to conduct a statistical analysis of significance, and potentially contribute thereby to the practice of palaeography, provided the results can be presented in a meaningful form.

In this work, we employ methods that are commonly used for mining insights from biological experiments regarding underlying genetic mechanisms. We show that in the context of palaeography such an approach also provides insightful observations.

Methods

A dataset consisting of 456 scribal handwritings in English Vernacular minuscule, ca. 990 – ca. 1035, is used (Stokes, 2005). "Scribal handwriting" here refers to a single stint or block of writing by one person; these samples are spread across some 198 manuscripts and range from the main text of the book to later additions and notes or glosses between the lines or in the margins.

The handwritings were described using 289 descriptors (Stokes, 2007-2008), where each descriptor indicates whether a certain grapheme (or group of similar graphemes) written as specific allograph(s) appear in the manuscript, as well as forms of certain parts of letters such as ascenders, descenders, and pen-angle. Every sample of handwriting is described by its known or predicted place of writing (where possible) and the estimated range of dates of writing.

For classification purposes, the dataset was divided into classes that are homogeneous in time and place.

First, we measure for each descriptor how informative it is. This is done using the *information gain* method (Mitchell,1997), which is often used for feature selection in text categorization tasks. The information gain score measures the decrease in entropy when a descriptor is given vs. the baseline in which it is absent. That is, the information gain measures the discriminative power added by each single descriptor.

Measuring the power of each descriptor by itself is of limited power. That a certain descriptor is ranked high tells us very little about how informative other similar descriptors are. Using an analogy to biology, it is helpful to know which genes express differently in a specific experiment testing different classes of biological conditions. However, if we want to learn about biological functions and processes, we must go beyond the level of the specific gene by aggregating the information from multiple genes of the same family. In palaeography, the importance of a specific character, for example, cannot be reliably detected just by looking at the ranking of one of the associated descriptors, no matter how highly it is ranked.

The Gene Set Enrichment Analysis (GSEA) is a statistically valid tool to evaluate how prominent a set (that is, a family) of descriptors is (Subramanian, 2005). This computational method determines whether an a priori defined set of features presents itself in a statistically significant and coherent manner among a ranked list of descriptors.

The input of the GSEA method is a ranked list of features and a list of families of features. Context is important, and so we conduct one experiment using families derived from graphemes and another experiment for families of allographs. The GSEA method is based on elaborate order statistic mechanisms that can compare, on a leveled ground, between large and small families of features.

Results

The information gain method was used to automatically rank all 289 descriptors. The first, most significant, 30 features are listed in Table 1.

Rank	**Descriptor**	**Rank**	**Descriptor**
1	æ__horned	16	7__high_top_right
2	a__horned	17	d__bilinear
3	s__round	18	g__convex_top
4	æ__angled tongue	19	g__tail_in_middle

5	d__vert_tipped	20	æ__rotund_minim
6	f__deep_split	21	y__hooked_tail
7	æ__high_e	22	æ__low_lig
8	æ__flat_topped	23	7__convex_top
9	g__tail_oblong	24	æ__tall_non_bulging_lig
10	a__flat_topped	25	e__angled_SW
11	æ__Caroline	26	a__angled_top
12	ð__vertical_tip	27	c__flat_hook
13	c__angled_SW	28	Asc.__forked_trailing_to_left
14	a__teardrop	29	Þð__predominant_ð
15	7__vertical_desc.	30	Aspct__rounded

Table 1: The 30 single descriptors that were found to be the most informative using the information gain score.

GSEA was then applied to these results, finding 16 families of graphemes (and stylistic issues, like aspect) that are “enriched” in a statistically significant way, and 20 families of “enriched” allographs. Tables 2 and 4 show the enriched families. The GSEA tool also provides another output: a list of descriptor families that are statistically speaking irrelevant. These lists (Tables 3 and 5) contain 5 graphemes and 9 allographs whose descriptors are ranked so consistently low that it is unlikely to be by chance. As can be seen, the GSEA results contain both expected results and surprising ones; see (Ker, 1957; Dumville,1988; 1993; 1994) for general background for script of this period.

Rank	**Grapheme+**	**Normalized ES**
1	æ	1.73
2	7 (Tironean nota)	1.31
3	a	1.31
4	c	1.15
5	f	1.10
6	d	0.86
7	g	0.82

8	Aspect	0.79
9	e	0.76
10	y	0.64
11	Ascender	0.60
12	h	0.58
13	Þð	0.58
14	k	0.57
15	Descender	0.56
16	s	0.50

Table 2: The 16 graphemes (and stylistic issues) that were found to be discriminative in a statistically significant way by the GSEA method. The enrichment score (ES) reflects the degree to which a feature-set is overrepresented at the top. The normalized ES accounts for differences in set size and the correlations between the sets.

It is unsurprising that the ash (æ) is most significant, since it is a combination of *a* and *e*, both of which are in themselves significant (Dumville, 1988). It is quite surprising that the Tironean nota for *and* (which looks like the numeral 7) is discriminating, as this is not noted by any of the palaeographers cited. It also came as a surprise that *c* and *y* turned out to be highly significant, because they are not generally recognised as such, though some of their specific forms certainly are (Dumville, 1993). It was also surprising, from a palaeographer's viewpoint, to see *h* in the list, as it is not included in those published by palaeographers (Ker,1957; Dumville, 1993; 1994; Stokes, 2005).

Rank	**Graphemes+**	**Normalized ES**
1	Minim	-1.20
2	Pen	-0.90
3	hmn	-0.76
4	r	-0.70
5	ð	-0.62

Table 3: The five graphemes and stylistic issues that were found to be irrelevant to the discrimination task to a statistically significant degree.

Here, it is very interesting to see that ð is insignificant, since its form varies very widely between scribes. Perhaps some of its features are better discriminants than others.

Rank	**Allograph**	**Normalized ES**
1	HORNED	1.63
2	CONVEX_TOP	1.51
3	FLAT_TOPPED	1.48
4	ANGLED_SW	1.32
5	ANGLED_TONGUE	1.30
6	HORIZ._TONGUE	1.15
7	TEARDROP	1.08
8	BILINEAR	1.08
9	SEMI_CAROLINE	1.07
10	ROUND	1.03
11	ANGLED_TOP	0.92
12	ANGLED_SHOULDER	0.91
13	LOW_LIG	0.90
14	ROTUND_MINIM	0.89
15	FLAT_HOOK	0.82
16	SHORT	0.82
17	LONG_TONGUE	0.76
18	CAROLINE	0.75
19	BULGING_LIG	0.74
20	SQUINTING	0.73

Table 4: The 20 allographs that were found to be discriminative in a statistically significant way by the GSEA method.

Table 4 is also quite insightful. While HORNED is widely accepted as significant (Ker, 1957), ANGLED_SW (angled southwest quadrant) is not, although it has been suggested as discriminative (Stokes, 2005). LONG_TONGUE is not attested in the literature as significant, as are the related ANGLED_TONGUE and HORIZ._TONGUE: it will be interesting to study if they are strongly correlated, and if they are even more distinctive in combination. Since TEARDROP, ROUND, and SEMI_CAROLINE are all forms of the letter *a*, their significance is expected and is supported by related literature on the minuscule of the period (Ker, 1957; Dumville, 1994), and is strongly argued as relevant for this period in the thesis from which this dataset is taken (Stokes, 2005). ROTUND_MINIM is unrecognized as being discriminative. LOW_LIG (low ligature) is less recognized in the literature than TALL_LIG (Ker, 1957) which is correlated with it. BULGING_LIG is well recognized (Ker, 1957; Stokes, 2005). SQUINTING is recognised as distinctive in Latin (Dumville, 1993) but not in the script normally used for the vernacular.

Rank	Allograph	Normalized ES
1	MINIM_LENGTH	-1.35
2	TALL	-1.05
3	MINIMS_ROUNDED	-1.04
4	LONG	-0.98
5	SHORT_HOOK	-0.89
6	TURNED_DOWN_TONGUE	-0.81
7	WEDGED	-0.78
8	ATTACK_STROKE	-0.66
9	STRAIGHT_BACKED	-0.58

Table 5: The nine allographs that were found to be irrelevant in a statistically significant manner to the discrimination task. Most of these results are expected since these allographs are either very common in the corpus or present the 'default' values for the script of the period.

Discussion

Computerized systems that perform digital palaeography have been criticized in the past for reducing script entirely to statistical processes that are themselves difficult or impossible to evaluate (Stokes, 2009). In fact, the struggle to elicit meaning out of statistical inference tools is common to many scientific domains. Here we begin to show that, by using the appropriate statistical tools, computers can be used to mine meaningful insights in palaeography.

The proposed method is not without its limitations. First, it relies on a specific definition of 'distinctive descriptors' that is derived from the choice of the feature ranking algorithm used. The IG method used focuses on the ability to discriminate between the classes; by choosing another ranking method one could focus, for example, on scribal variation, which is also a question of interest to palaeographers. A second limitation is that the method cannot go beyond the assumptions made in the initial coding system. For example, there is a normalization quality to GSEA, in the sense that, if the palaeographer recorded more varieties of *a,* say, than of other letters, precisely because he expects that letter to be more significant, the GSEA method will counterbalance this by looking at the overall distribution of all varieties. However, presumably there could have been other letters and features that are in fact more distinctive but that were not recorded in the database because they were mistakenly deemed to be relatively insignificant. Future work should therefore aim to augment the database with automatically extracted features, with the potential benefit of adding a new (robotic) perspective to morphological analysis. Another limitation of the method is its dependence on verbal descriptors for features which are visual, or, indeed, which are a function of the physical movements of the scribe's arm, hand and pen, particularly given the lack of standard palaeographical terminology for such detailed features (Stokes 2011–12). These difficulties could potentially be overcome by providing greater rigour in nomenclature and by connecting these labels to particular images. Both approaches are already being tested in the DigiPal project (http://digipal.eu), and discussions are already underway to combine the work done there with that described here.

Despite these limitations, the method described is still very promising. One of the difficulties in palaeographical study is the vast quantity of detail that must be processed, and so helping the human expert to identify distinctive features would be enormously beneficial in managing that data. For instance, identifying patterns in variation such as those by region, date or group of scribes would be invaluable in identifying manuscripts by their writing. Applying the method to other corpora of scribal handwriting could also lead to valuable insights. For example, if particular features prove to be discriminative across many different script-systems then this could be an important clue into identifying scribes independently of the script that they wrote, something very necessary given that most scribes routinely wrote in a number of very different scripts. Finally, with further testing and refinement, this approach promises an important and large step towards a method for distinguishing handwriting that can be described explicitly, that can be communicated effectively in ways that palaeographers and other humanities scholars can understand, and that can also enjoy the support of quantitative data. This goal is one that has been sought for a very long time (Stokes 2009).

References

Bishop, T.A.M. (1971). *English Caroline Minuscule*. Oxford: Clarendon Press

Dumville, D.N. (1988). Beowulf come lately: some notes on the palaeography of the Nowell Codex. *Archiv für das Studium der neueren Sprachen und Literaturen*, **225**: 49–63.

Dumville, D.N. (1993). *English Caroline Script and Monastic History*. Boydell: Woodbridge.

Dumville, D.N. (1994). English Square Minuscule script: the mid-century phases. *Anglo-Saxon England*, **23**: 133–64

Ker, N.R. (1957). *Catalogue of Manuscripts Containing Anglo-Saxon*. Oxford: Clarendon Press

Mitchell, T.M. (1997). *Machine Learning*. New York: McGraw-Hill

Stokes, P.A. (2005). *English Vernacular Script ca 990–ca 1035*. Cambridge [unpublished Ph.D. dissertation].

Stokes, P.A. (2007–2008). Palaeography and image-processing: some solutions and problems. *Digital Medievalist* **3**. http://digitalmedievalist.org/journal/3/stokes/ (accessed 15 March 2012).

Stokes, P.A. (2009). Computer-aided palaeography: present and future. In Rehbein, M., *et al*. (eds), *Kodikologie und Paläographie im Digitalen Zeitalter*. Norderstedt: Books on Demand. 309–338. http://kups.ub.uni-koeln.de/2978/ (accessed 15 March 2012).

Stokes, P.A. (2011–12). Describing script [Parts i–]. *DigiPal Project Blog*. King's College London. http://digipal.eu/blogs/blog/describing-handwriting-part-i/ (accessed 15 March 2012).

Subramanian, A. *et al*. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(43): 15545–50.