University of Lund
Department of Linguistics

Allograph-Based Categorization of Handwriting Styles

Nils Rosengren

BA thesis in General Linguistics /
C-uppsats i allmän språkvetenskap
May 2002

Tutors Mats Eeg-Olofsson, Department of Linguistics
Magnus Nordenhake, Decuma AB

Abstract
The aim of this work is to aid the development of online handwriting recognition software by using a linguistic approach to the problem of categorizing handwriting styles. This is done by exploring the wide variety of styles among users of the Latin alphabet, trying to explain the variety and finding ways of categorizing the styles. The variety is due to the many variables open to the writer and the wide range of these variables. A generative model for the production of hndwriting is proposed. Different ways of categorizing styles along the lines of the variables are explored, and one of them, allographic variation, i.e. different models of the same letter, is selected for the study of a database of online samples consisting of handwritten block letters from six European countries. Not only the shapes of letters are examined but also the number and sequence of the constituting penstrokes, which is especially interesting from an online recognition point of view. It is found that there are different 'alphabets' or categories of allographs, as well as geographical differences in the adherence to the alphabets.

Contents

1. Background 3
1.1. Graphonomics and Handwriting Recognition 3
1.1.1. What is HWR? 3
1.2. The variety of handwriting 3

2. Aspects of writing 4
2.1. Some graphonomic theory 4
2.2. Constituents of handwriting 4
2.2.1. Arcs and strokes 4
2.2.2. Sequencing 4
2.3. Block letters vs. cursive 5

3. Some questions of style 5
3.1 Forensic science and handwriting styles 5
3.2. HWR research: efforts at categorization 7

4. What happens between brain and paper? 9
4.1. Sources of variety 9
4.2. Manifestations of variety 10
4.3. A generative model for the production of handwriting 10

5. Research and method 11
5.1. Material 11
5.2. Allograph-based style categorization 12
5.3. The allograph 13
5.3.1. Visible allographic variety 13
5.3.2. Invisible allographic variety 13
5.3.3. Allograph criteria 13
5.4. Hypothesis: The personal alphabet 13
5.4.1. Pervading characteristics of the personal alphabet 13
5.5. Sorting allographs 14
5.6. Looking for patterns 14
5.6.1. Upper case allograph groups 18
5.6.2. Lower case allograph groups 19

6. Results and analysis 19
6.1. Allograph group adherence 19
6.2. Geographical distribution 21
6.3. Other factors 21
6.3.1. Number of allographs as a measure for neatness 21

7. Discussion 22
7.1. The Latin and Germanic style groups 22
7.2. Two extremes: France and Sweden 22

8. Conclusion 22

9. References 23

10. Swedish equivalents of some English terms 23

1. Background

1.1. Graphonomics and Handwriting Recognition

The term *graphonomy* for the linguistic study of writing was coined in the 1960's. The name was chosen to distinguish it from graphology, to which it perhaps has the same relation as astronomy to astrology. Within the field of general linguistics, it could be seen as the lesser-known and ill-fated equivalent of phonetics, as it studies writing the same way as phonetics studies speech. It was introduced in Sweden by Allén (1971). After falling into oblivion it was revived in 1982 with the founding of the International Graphonomics Society. An interdisciplinary research field, *graphonomics* is aimed at studying all phenomena connected to handwriting.

As pointed out by Sampson (1985), there has been little interest within the field of linguistics for the study of writing as such: the interest has mostly been turned towards the connection between speech and writing. Over the last few decades, a new need for handwriting-related research has evolved with the evolution of computerized handwriting recognition systems (for an overview of the field, see Plamondon & Srihari 2000); this has also given new opportunities of research by making an *online* study of handwriting possible. With its experience of developing models for linguistic behaviour as well as of the statistical handling of language data, linguistics might do a lot of good here.

1.1.1. What is HWR?

Handwriting recognition technology (HWR) means converting handwritten letters to a digital representation (the 8-bit ASCII-code used in computers). To put it quite simply, the keyboard is replaced with the pen as a means of computer input. This would be of great interest to the cultures of East Asia with their multitude of characters, but the Latin alphabet users of the Western market are also very much considered.

There are two kinds of handwriting recognition: offline HWR, found in e.g. post-sorting systems, where a complete text is scanned and converted, in a process similar to human reading, and online HWR, found in hand-held computers, where the text is recognized while written, by having the compter sensing the movements of the pen. This study is aimed at online-HWR. Although this technology has improved much in recent years, the step from the developer's lab to real-life use of the product is still difficult.

1.2. The variety of handwriting

No two instances of the same handwritten letter look exactly alike. Just as any other product of human activity, handwriting shows great *variation* and *variability.* With *variation* we mean the fact that no two persons write in the same way: the differences in handwriting between different people. With *variability* we mean that the handwriting of a given person varies over time. The variability is generally explained by 'personal, emotional, intentional, circumstantial factors' (Crettez 1995). Still, it is reasonable to assume that the force of variation is greater than that of variability: no matter how much the handwriting of a given person varies, it will still retain the qualities that distinguish it from that of other people. To talk about both these forces at once, we will use the term *variety*.

All this means that as well as conveying a linguistic message to the reader, a piece of handwriting contains a lot of 'noise', information originating from and personal to the writer. All HWR-systems need to filter away this noise to get at the level of meaning. The difficulty of this task lies in the unpredictability of the variety. This is the main reason why no HWR-system as yet succeeds in recognizing 100% of the input. According to Dutch researchers Vuurpijl and Schomaker (1997), the recognition rates reported by academic and commercial developers are usually overestimated by 10-20%. This is explained by the fact that although it

should be possible to 'train' any good HWR system by exposing it to a wide variety of handwriting styles, it is still unprepared for the kinds of messy, speedy handwriting found in the real world. This study is an investigation into the nature of handwriting and the variety of handwriting styles. We will see whether a simple, manual method is enough to establish a first level of categorization in the variety.

2. Aspects of writing

2.1. Some graphonomic theory

Graphonomic theory adopts the same *generative* view of writing as phonetics does of speech. It is assumed that all linguistic activity is founded upon an 'underlying form', connected with rules to the visible 'surface form'. In speech, the underlying form is the phoneme, in writing it is the *grapheme*, or the 'inner image' we possess of a letter. This means that the letter does not exist *as such* in the real world: there are only different *instances* of a grapheme. A single, handwritten character we call *graph*. In fact, a graph could be any sort of scribble; it is up to the reader to interpret it, i.e. assign it to a class of graphemes. In this infinite variety no one could claim to have found the perfect, archetypical letter. The grapheme can be a letter, a number or a punctuation mark (respectively, alphabetical, numerical and junctural graphemes). This study only concerns alphabetical graphemes. In graphonomic notation, graphemes are written between angles: <g>.

In one end, the grapheme is associated with one or many phonemes according to language-specific rules (a relationship analysed in Allén 1971). In the other end, the grapheme is associated with the specific movement pattern, called *ductus* by Plamondon & Srihari (2000), needed to trace it on a surface. For each grapheme, several movement patterns are possible. This gives rise to the existence of *allographs*, the equivalent of the allophones in phonetics. As typical specimens of allographs, Sampson (1985) mentions <g> and <g>: two clearly different versions of the same grapheme used in different fonts. In handwriting the situation is somewhat more complex: this will be investigated below. One could imagine that the number of allographs rises with the complexity of the grapheme, and that a grapheme might be associated with more than one allograph by a given writer.

2.2. Constituents of handwriting

2.2.1. Arcs and strokes

We need to go into somewhat meticulous detail about the way handwriting is done. A letter, number or punctuation mark, in short a *graph*, is made with one or more *arcs*, a term coined at Decuma AB. An arc is the track of the pen separated by two 'lifts'. During the arc, the pen can change direction, causing *strokes*. For example, the letter <L> is normally produced with one arc consisting of two strokes. An arc has a startpoint, which can be placed at any of its ends, and an endpoint. Consequently, the letter <I> can be drawn in two ways: top to bottom or bottom to top.

2.2.2. Sequencing

Furthermore, the arcs may be drawn in different orders: you can draw the letter <T> either by beginning with the horizontal or the vertical line. By varying the startpoints, you get two more varieties. However, a more complex letter such as <E>, if produced in four arcs, each of which can be started at two ends, may be drawn in no less than 384 different ways.

None of this is visible when looking on the finished letter. But it poses problems for online HWR. All possible starting points and arc sequences need to be taken into

consideration. Another interesting consequence is that online HWR makes it possible to record and store handwriting electronically. You can replay it to study startpoints, number of arcs and sequencing. This method is used in the present study.

2.3. Block letters vs. cursive

There are two kinds of handwriting: block letters, where the letters are written separate from each other, and cursive, where the letters are connected, the ideal being that a word should be written with a single arc. No further distinction is here made as to the 'naturalness' or 'personality' of the two systems: as we shall see, block letters can be as peronal as cursive script. Practical experience tells us that most people write in a mix between the two, if they are at all able to make out the difference.

The ideal of HWR developers is of course that the user should be able to write as naturally and unimpededly as possible. Today, though, the users are restricted to writing block letters. The problem of having the recognizers distinguish the letters of cursive script remains to be solved. Perhaps a thorough investigation of block letters is needed before cursive script is attacked: this study is restricted to block letters.

3. Some questions of style

What, exactly, is a handwriting style? Is there something more to the question than popular conceptions such as that girls write more neatly than boys, that doctors write illegibly, that people used to write more legibly and beautifully in the old days, etc? Why doesn't everybody write in the same way? We need to get behind some popular conceptions here.

3.1 Forensic science and handwriting styles

What is meant by 'style'? A vague and wide term, it can be defined as 'a set of pervading characteristics'. In colloquial speech, 'writing style' might mean the particular style or 'hand' of an individual. It could also apply to a characteristic trait shared by a group of writers: we talk of messy, upright, dense writing styles and so forth. The different kinds of writing learnt at school are also called writing styles.

Following British forensics scientists Brown and Davis (Davis 1994), I would like to call the latter not styles but *systems* of handwriting. In their 1982-83 investigation, they tried to map out the teaching of handwriting in the United Kingdom, and to investigate the relation between this education and the way handwriting is actually done by grownups. It was found that four different systems of handwriting were taught at schools (Figure 1): three kinds of cursive writing, 'Looped Cursive', 'Round Hand' and 'Italic', and one kind of block letters called 'Print Script'. The systems are all results of different notions of beauty and expediency, differing both as to the shape of the letters and the way they are (or aren't) connected. The geographical distribution of the systems is complex but not chaotic; according to Davis, British handwriting is more various than in other European countries, where an official handwriting system often is imposed by the authorities.

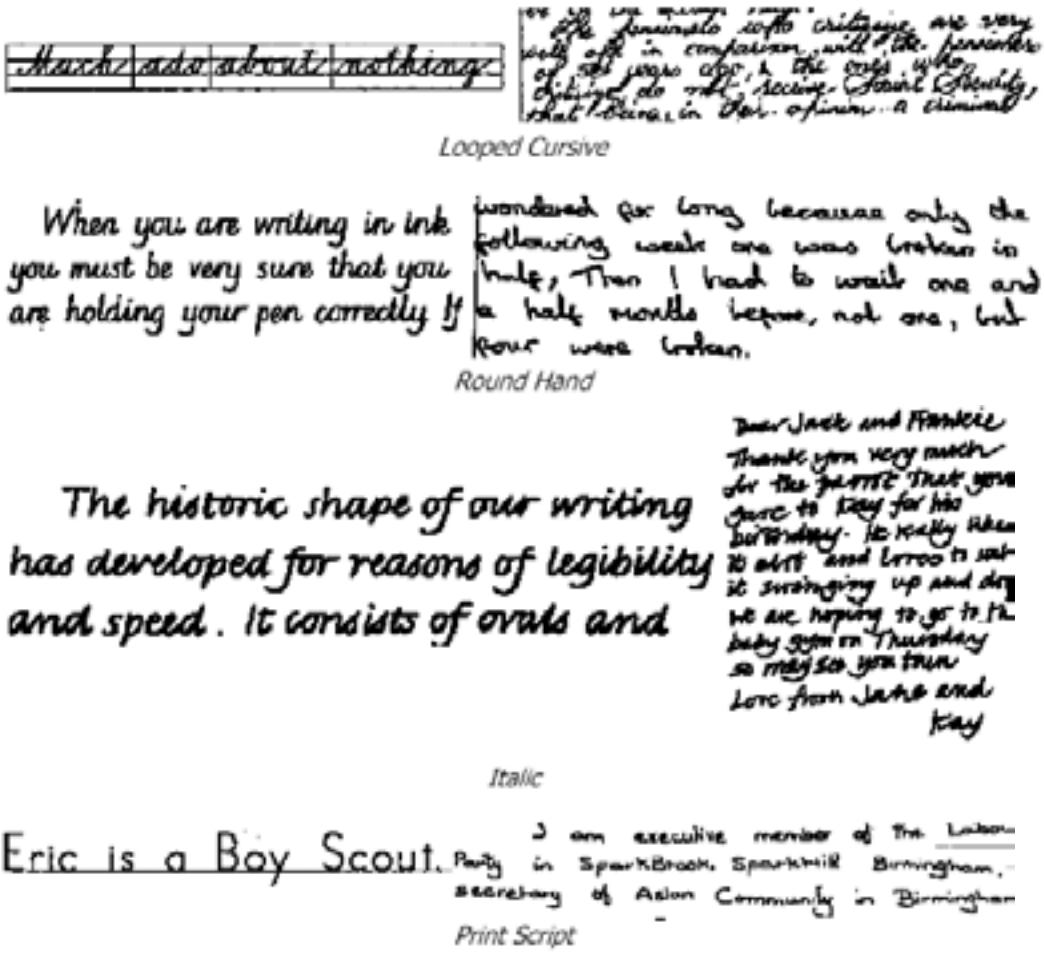


Figure 1. Four British handwriting systems (after Davis 1994). To the left are copybook samples, to the right are 'real' samples from adult writers

Of course, nobody really writes as in the copybook examples of Figure 1. If this was the case, HWR-systems would have a recognition rate of 100%. What is more, nobody is expected to slavishly copy the systems, but rather to develop their own idiosyncratic 'hand'. This process could be summarized as follows. Learning to write, a group of people take in a certain handwriting system to different extents. While growing up, they more or less deliberately modify this into an individual style, through the establishing of different characteristics.

These characteristics Davis separates into *individual characteristics*, 'those components of a given hand that make it unique', and *style characteristics*, 'those features which are shared with other members of a group or groups'. Such a group I would like to call a *style category*. It might be argued, however, that there is no such thing as individual characteristics, but rather that it is its specific set of style characteristics that makes a certain hand unique. A handwriting style, then, I define as that set of characteristics.

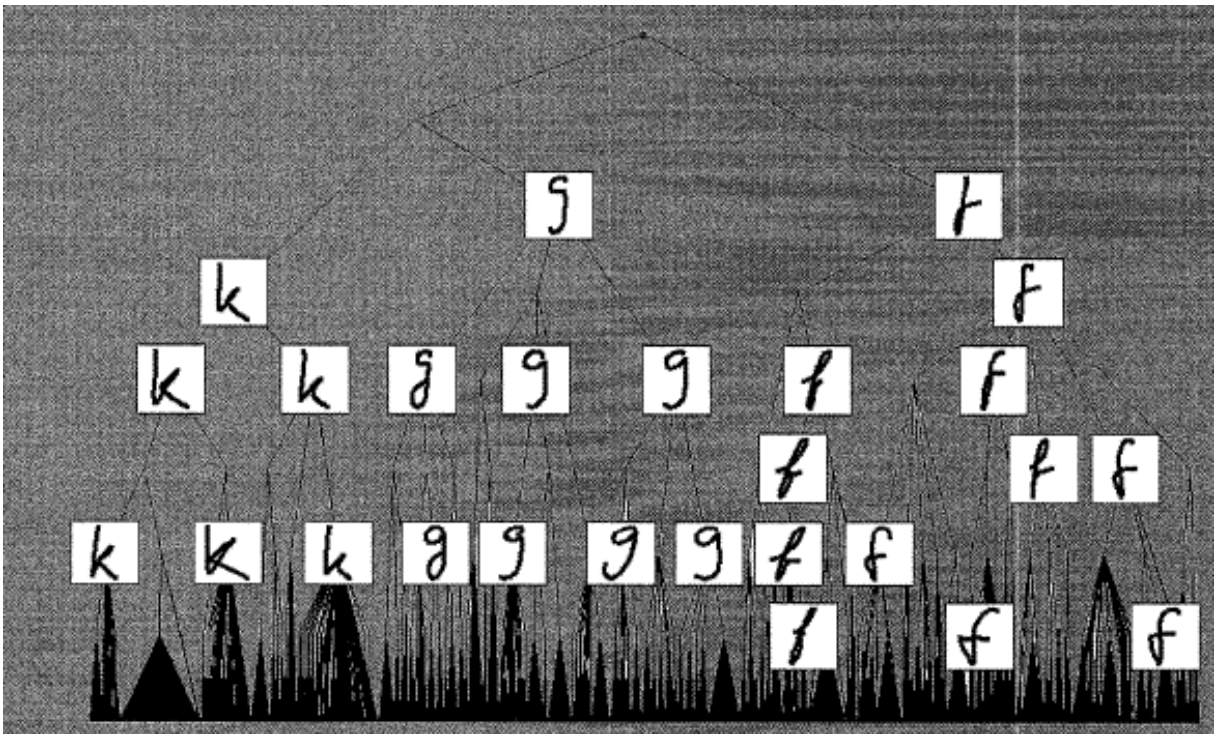


Figure 2. Hierarchical clustering of allographs (after Schomaker 1998). A graph within a box is the average of all members contained therein

3.2. HWR research: efforts at categorization

Until now, the science most interested in handwriting styles has perhaps been criminology. Forensic document examiners try to attribute a handwritten document to a particular person and have developed considerable skill. There is of course also graphology, a pseudo-science that tries to draw conclusions on the personality of the writer from the nature of the style. Lately, HWR-experts have turned their interest to writing styles. Vuurpijl and Schomaker (1997) claim that there is no catalogue of Western handwriting-styles; they also point out that explicit knowledge of different forms of handwriting might increase recognition rates.

The goal is to see whether there are differences in style between nationalities, sexes, ages and social groups. It is also interesting to see whether there are style categories independent of these factors. All this should help the HWR-systems to draw conclusions about the shape of the letters given facts about the writers. This would increase recognition rates, and could also be used for purposes of identification and security. However, the work done by these researchers has mostly consisted in collecting large handwriting databases and exploring the variety of styles; only recently one has felt the need to explore the mechanisms that lie behind the variation.

The HWR-researchers interested in the categorization of handwriting styles have gathered large corpuses of online handwriting samples in projects such as UNIPEN (described in Schomaker 1998). Using neural networks and other hierarchical clustering methods (Bote-Lorenzo et al. 2001, Schomaker et al 1994, Schomaker & Vuurpijl 1997) the graphs are sorted into clusters. When a cluster is found, it is made to represent one prototypical allograph (Figure 2). The allograph extraction methods have been shown to improve the performance of recognizers. The basis of the allograph categorization is the concept of the *stroke,* defined as

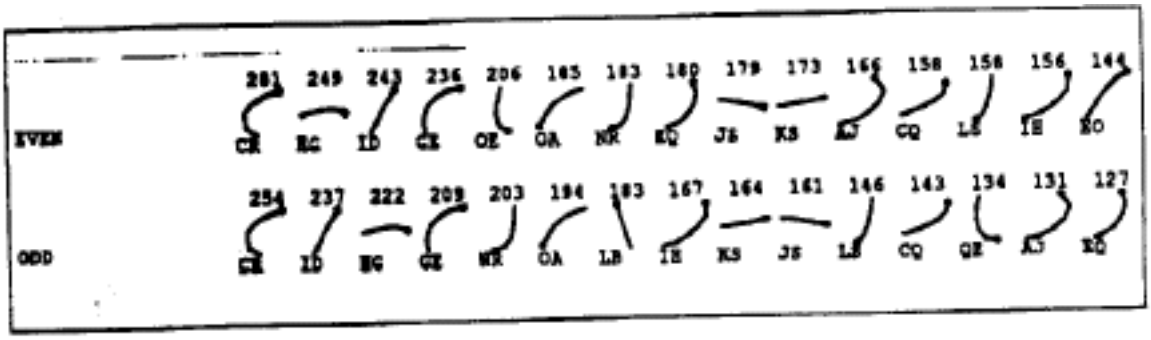


Figure 3. The fifteen most common strokes penstrokes from two sets of handwriting samples from the same writer (after Schomaker et al. 1994). As we can see, they are about the same in both sets

the smallest unit of pen-tip movement. The segmentation into strokes is made in different ways (see Plamondon & Srihari 2000): at maximal points of curvature, at points of minimal velocity, at minima of the *y(t)* function (i.e. lowest points of graphs), etc. An interesting result of these inquiries are the 'prototypical strokes' found by Schomaker et al. (1994) and their connection to nationality and sex (Figure 3).

Other interesting results have been found by Crettez (1995). Starting from the fact that writers draw their strokes in some directions more frequently than in others, which decides the general slant and spread of the handwriting, he finds that each writer generally has four 'preferential directions'. The directions are used to establish a number of handwriting 'families' that can be used for adapting recognizers (figure 4).

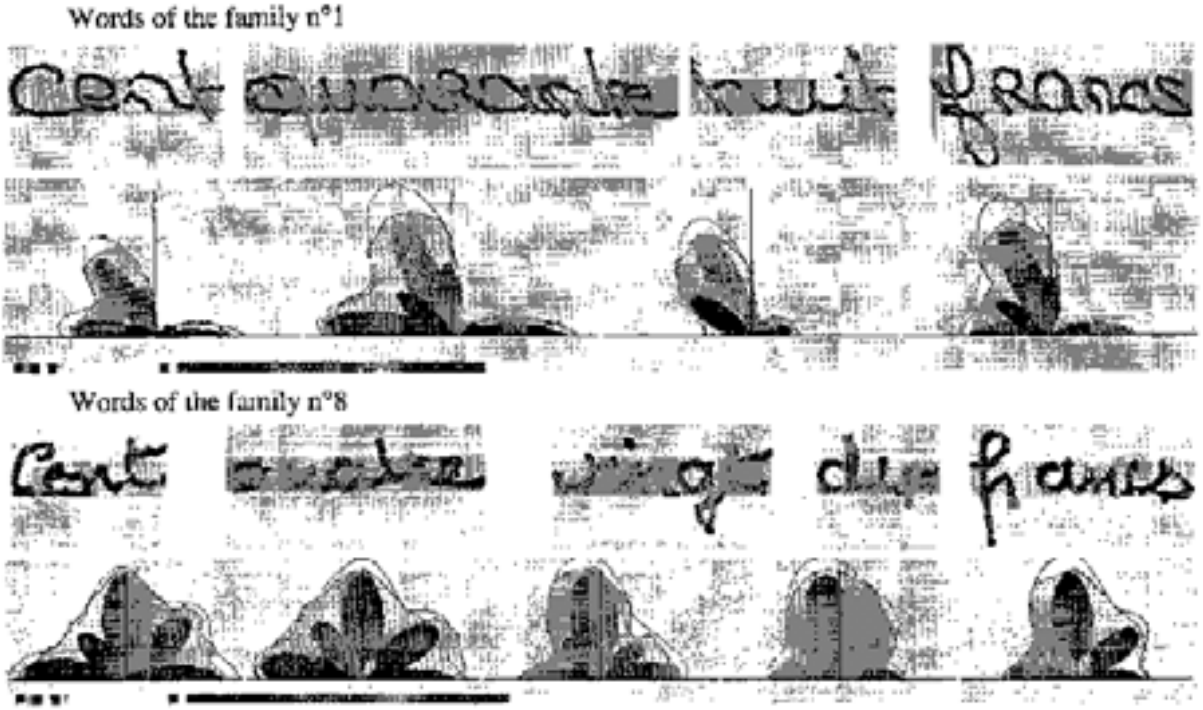


Figure 4. Two handwriting families, based on preferential directions (after Crettez 1995). Below each word is its directional diagram; words of the same family have about the same diagram

4. What happens between brain and paper?

The Latin alphabet consists of 25-30 different characters (depending on the language) which over the years have developed into easily recognizable shapes (for an outline, see Gelb 1963). How could it be that these well-known models should produce such different results? The process of handwriting is something like the opposite of online-HWR: the representation of a letter in the brain is converted to the muscular movements needed to draw the letter on the surface. While travelling the distance between brain and paper the letter has to go through a lot. Without going into too much anatomical detail, we might say that on its way through the 'writing apparatus', the signal passes through the cortex and the nerves; the movement is transmitted through limbs, joints and muscles, all with their own dimensions in every person.

4.1. Sources of variety

Although the situation may seem messy and complex, it is quite possible to point at the different factors involved. Schomaker (1998) identifies four sources of variety in handwriting. We will adopt these with some modification (Figure 5).

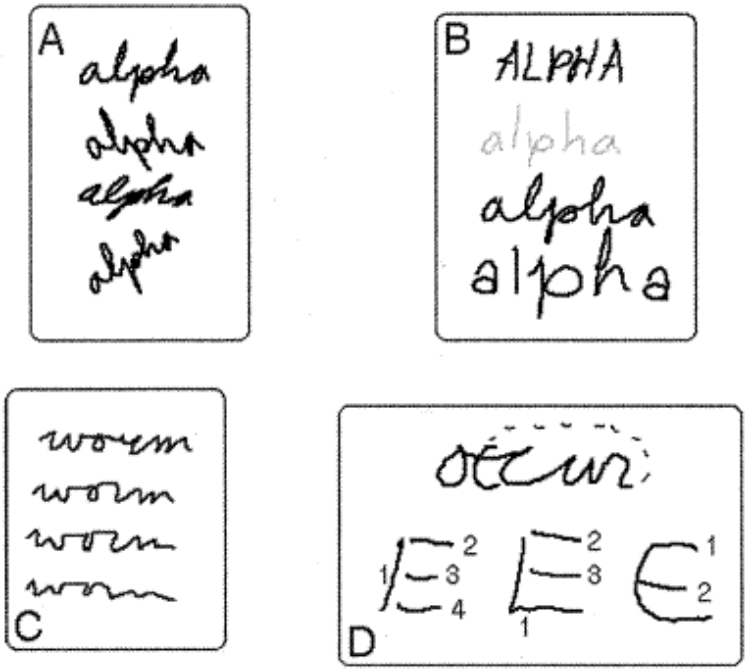


Figure 5. Four sources of variety in handwriting (after Schomaker 1998). A. Affine transforms B. Allographic variety C. Neuro-biomechanical variety D. Sequence variety

A. Affine transformation A geometrical term, meaning that a figure can be changed in different ways while retaining the same basic shape. In handwriting, this means that size, slant, etc. can be varied by the writer.

B. Allographic variety Each grapheme is associated with a certain model, or allograph, by each writer. A writer may use several allographs for one grapheme.

C. Neuro-biomechanical variety The neurological and biomechanical properties of the writing apparatus of different people give rise to differences. With the same writer, the writing can vary depending on concentration and speed.

D. Sequence variety The number and order of arcs in the graphemes may vary. Writers may turn back to add forgotten letters and diacritic marks (for example, dotting your i's after having written a whole word).

Further examples can be found in examining Figure 1.

A. Compare the different instances of <g> in the Round Hand sample, or the <B>'s in Print Script.

B. Compare the <S> in "Social Security" (Looped Cursive) with <S> in "Sparkbrook, Sparkhill" (Print Script), or the <y>'s in the copy-book samples of Italic and Print Script.

C. Differences in heigth, width and slant between all four samples.

D. Difficult to discern in completed text, this variation is visible in online-recored handwriting. See also [5.3.2. Invisible allographic variety].

Another source of variety is the fact that about one-eighth of all people are left-handed and might be expected to use their own allographs of virtually all graphemes.

In handwriting recognition, the toughest problem is posed by types (B, D), because of the many possible allographs and sequences. For reasons explained later on, I would like to unite these factors into one, to use it as the basis of a categorization.

4.2. Manifestations of variety

As pointed out by Crettez, you can make some manifest observations appliable in categorizing the writer's style. Apart from the allotraits (directions of tracing more common than others), there are the thickness of the tracing, the size of the main body of the writing compared to that of descenders and stems, and the spatial density of the characters.

It is the combination of these variables that give rise to the variety of styles and the unicity of each style. Through exploring the variables involved and by placing each style at its place along an axis in a hypercube, it would be possible to categorize each style. For this kind of categorization, computer-aided methods are very well suited. But this is beyond our scope at the moment.

4.3. A generative model for the production of handwriting

At this point, I would like to construct a *generative* model of the handwriting process (Figure 5). This will help us further in understanding what happens from brain to paper. We suppose that not only the grapheme is layered in the brain but also its associated ductus. When writing a letter, the needed movement pattern is chosen: this is where allographic variation occurs. The 'prototypical strokes' mentioned in [3.2] might come in here as components of the movement pattern.

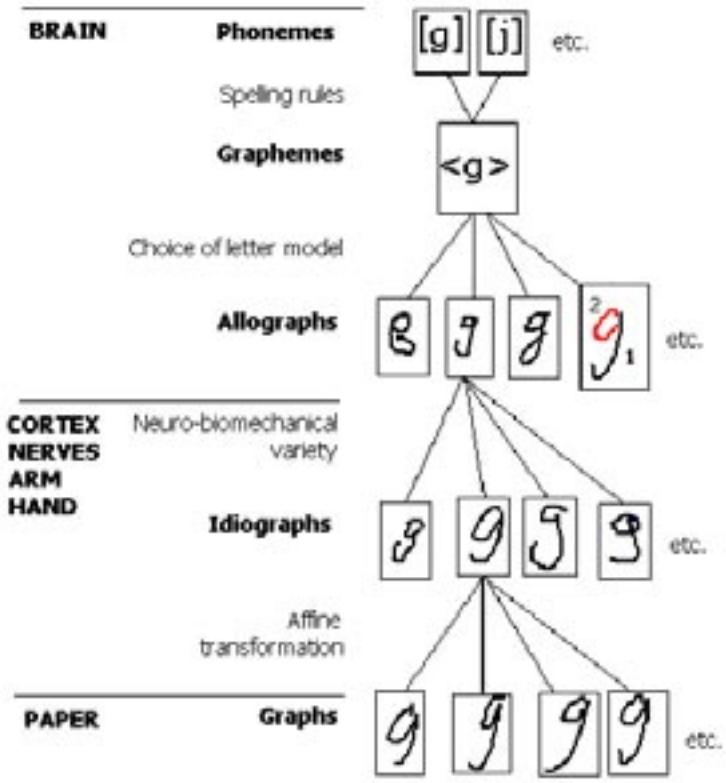


Figure 5. Generative model of the handwriting process from brain to paper

On its way through the writing apparatus, the letter undergoes neuro-motorical transformation to form the writer's personal variant of the grapheme. This is what we call the *idiograph*, and it could be seen as the 'average' of all instances of the same allograph produced by the writer. The four graphs on the idiograph level in Figure 5 are taken from different writers. - A possible objection against the idiograph concept might be that the idiograph does not exist as such. But as we remember this is not the case with the grapheme either.

The idiograph then undergoes affine transformation to produce the final appearance of the graph. The graphs on the bottom level of Figure 5 are all taken from the same writer. As we can see, they are clearly different yet somehow more alike each other than the idiographs above them.

5. Research and method

5.1. Material

During 2001, Decuma AB, a Lund-based HWR development company, has in a project similar to UNIPEN gathered a corpus of handwriting samples in Europe and North America. The aim was to study the variety of handwriting. The corpus consists of Latin block letters. The samples studied here were gathered in Germany (DE), Spain (ES), Finland (FI), France (FR), the UK (GB) and Sweden (SE). The samples were written with a stylus on the sensitive screen of a hand-held computer equipped with HWR. In order to guarantee block letters, the text was written in so called box mode, a format similar to an adress form, in which each letter is written in a box.

There are about 24 test persons (TP) from each country. Each one of them had two write a number of tasks: combinations of letters, numbers and punctuation marks. The TP's are

anonymous, but had to state their sex, age, level of education (on a subjective scale) and whether right- or lefthanded.

From each country I chose six TP's, making in all 36. The material is of course very small; still, it might give a sample of the allographic variety and indicate certain trends. To sort out one variable, only right-handed writers were chosen. The TP's are stored in the database in no particular order: to get a random selection I simply chose the first six persons in each country's catalogue. To get a I studied the around 40 tasks/TP that consists of real words, chosen so that each letter of that country's alphabet was represented (Figure 6). Each word was written in one lower case and one upper case version. The most common graphemes of course appear many times and the rare ones appear once or twice. The graphs are stored so that you can see them 'grow' arc for arc in a special program (SABED). The startpoint of each arc is marked with a dot.

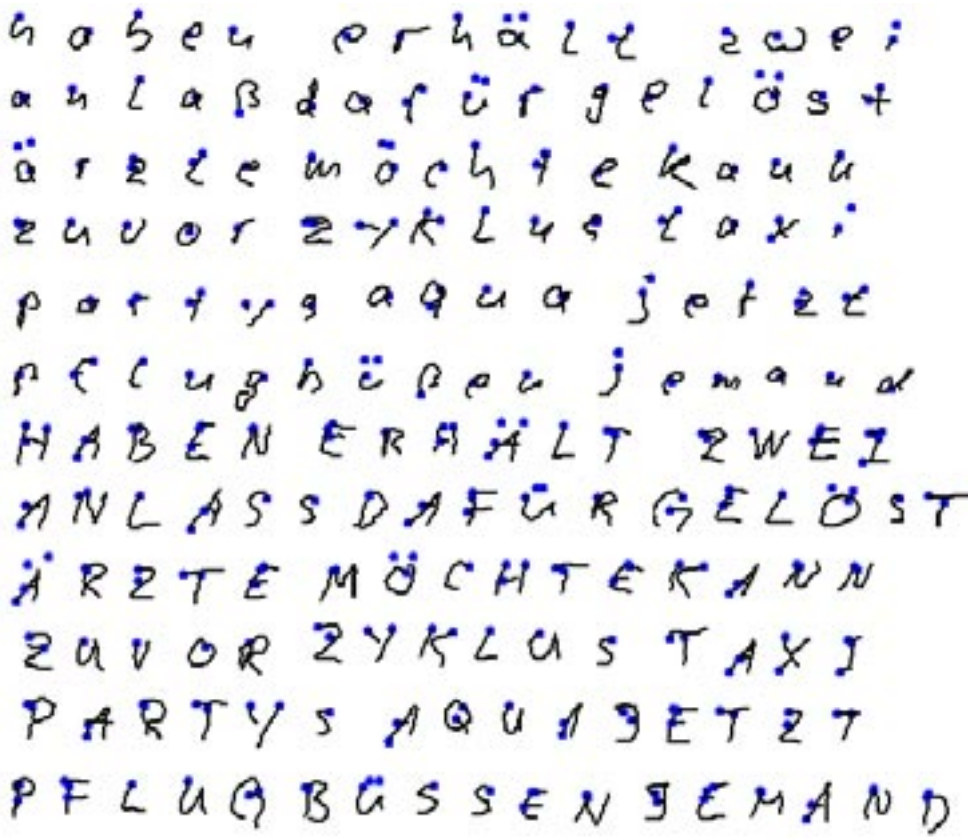


Figure 6. The words written by one test person (DE_03). Startpoints are marked with dots

5.2. Allograph-based style categorization

It now stands clear that there are not one but many ways of categorizing handwriting styles. Any of he types of variety can be used for establishing style characteristics shared by different groups of writers. It's all a matter about where you want to draw the line.

The heavy computer-aided methods mentioned in [3.2.] I found beyond my competence and resources. They are well adapted to large databases of graphs and to find variety dependent on neuromotorical and geometrical factors. The corpuses feature cursive handwriting or cursive/print mixtures. In the present study, we will see what can be done with a comparatively small database of block letters. One of the aims is to find geographical patterns in the handwriting of Europe. There is no reason to look for neuromotorical variety, since this would presuppose that different nationalities have different neuromotorical faculties. Instead, we will place the categorization higher up in the tree of Figure 5, at the

level of allographic variety. The arbitrariness of the stroke concept will make us use the arc as the foundation of allograph categorization.

5.3. The allograph

Since it's so difficult to find two handwritten letters that look the same, each graph could be said to represent its own allograph. We need to draw the limit earlier than that. We will establish allograph criteria and thus come up with a reasonable number of allograph prototypes used for classifying graphs. It's all about disregarding variation in letter shape emanating from the Schomaker (A, C) sources of variety, and finding the variation that really depends on different allographic representations of the grapheme.

Figure 7. Visible and invisible allographic variety

5.3.1. Visible allographic variety

From our corpus of handwriting samples we pick two instances of the grapheme <h>, written by the same person (Figure 7 A I-II). Do these represent different allographs of <h>? They differ visibly in shape but the differences are rather due to affine transformation.

From another writer, we pick two other instances of <h> (B I-II). Whereas B II clearly represents another allograph from both B I and A I-II, I would (for the time being) like to categorize BI as belonging to the same allograph as A I-II. Even if the arc is connected to the stem further down than in A I-II, the overall pattern is the same. Furthermore, apparently a writer might use more than one allograph for the same grapheme.

5.3.2. Invisible allographic variety

As shown in [2.2.2. Sequencing], different allographic representations do not necessarily give rise to visible differences between graphs, since you can't distinguish the number of arcs, their sequence or their starting points. By looking on the two graphs in A in SABED, we in fact find that they are drawn with two arcs, beginning with the stem (C I). The dots mark the startpoints of each arc; the latest-drawn arc is coloured grey. The two graphs in B, on the other hand, are drawn with one arc only. We at last get three different allographs of <h> (C I-III).

5.3.3. Allograph criteria

The allograph concept, i.e. variety that depends on different movement patterns, is expanded to include sequential variation: differences in number of arcs, arc sequence and arc starting points.

5.4. Hypothesis: The personal alphabet

We are now able to form a hypothesis. Apart from the style characteristics mentioned in [4.2. Manifestations of variety], each writer uses a set of allographs. We will see whether this 'personal alphabet' is arbitrarily composed or whether it has any pervading characteristics. Having done that, we will compare the alphabets of different writers, looking for larger groups of writers using similar alphabets.

5.4.1. Pervading characteristics of the personal alphabet

It might be expected that similar graphemes are executed in similar ways. For example, a writer using the two-arc allograph of Figure 7 C I of <h>, might also use two arcs for

executing similar graphemes such as <b, d, k, p, q>. The loop of Figure 7:B:II might also be found in <b, k> and in the descending stems of <g, j, y>. These were the kinds of connections expected when making the study, but we shall see that there are other, less apparent patterns.

5.5. Sorting allographs

In a pilot study, one writer from each country was examined. The aim was to form an image of the allographic variety: do all graphemes exhibit allographic variety, or can you disregard some of them? Furthermore: how much allographic variety does a writer exhibit, and how many allographs are used for each grapheme? It was also interesting to see whether any connections were visible between the allographs used by one writer, and whether any common traits could be found between such a small number of writers.

The files of all six TP's were went through, graph for graph. The graphs were sorted alphabetically and then after allographic variety, according to the criteria in [5.3.3]. In order that each test person's alphabet should consist of the same graphemes, the German <ß> and the French <Œ> were disregarded. The dotted graphemes <ü, å, ä, ö> occuring in German, Swedish and Finnish files were sorted under their dotless equivalents. For the 26 graphemes, 47 lower-case and 58 upper-case allographs were found.

Allographic variety was found for all of the graphemes except <b, c, e, h, m, n, o, s, w, C, L, O, Q, S, T, V, W>. This was unexpected for more complex graphemes such as <b, h, m, n, L, T>. It was decided not to disregard any graphemes in the main study. All TP's exhibited allographic variety for at least one grapheme. As an example, let me mention the British writer who used three different allographs of <f> in the same word, 'fluffy'. They are the ones markes f_1_1, f_1_4 and f_2_2 in Table 1.

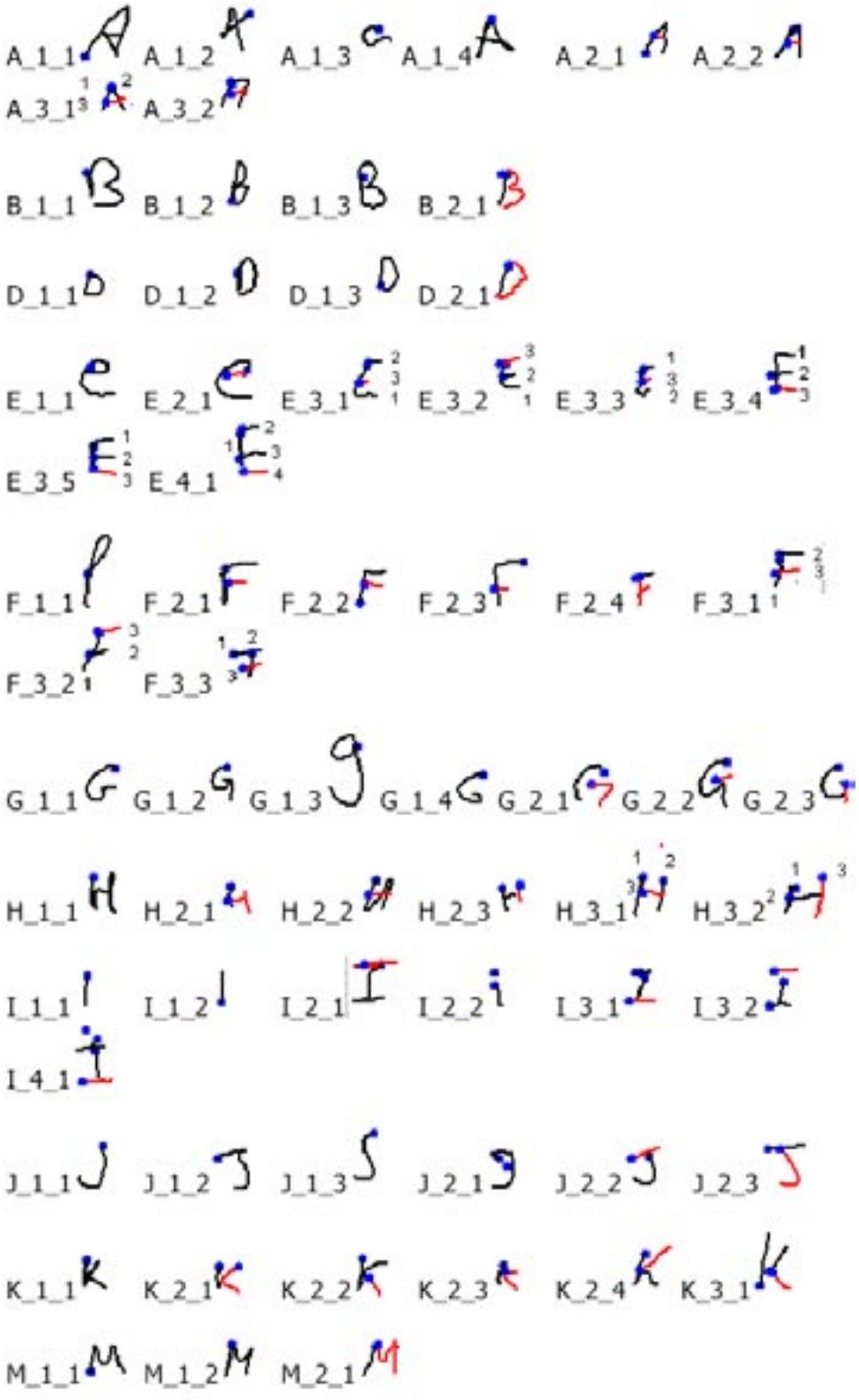
No patterns connected to nationality, sex or age were found. On the other hand, there were some observations of connections between the graphemes. This holds true especially for some of the upper-case graphemes, namely <A, B, D, F, K, M, N, P, R>. These patterns were fortified in the main study, in which the remaining files were sorted. After going through the remaining 30 test person's tasks (quite a time-consuming piece of work) the allograph prototypes established in the pilot study were extended with another 46 upper and 106 lower case ones (Table 1).

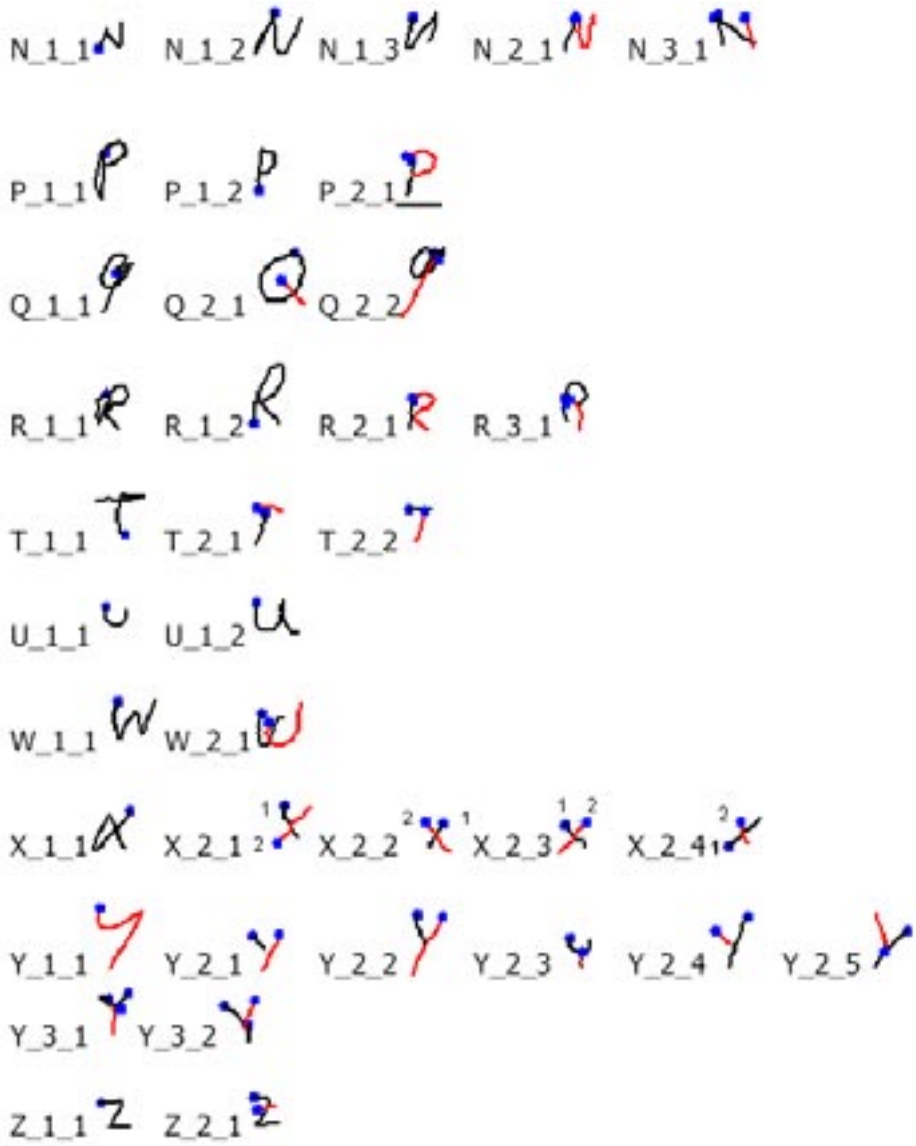
I will try to clarify the method used in categorizing the allographs by inviting the reader to take a look at the allographs of <A> in Table 1. Eight prototypes were found; each instance of <A> found in the material belongs to one of these categories. They represent, if you will, eight different ways of solving the problem of drawing an <A>. This letter consists of two oblique bars, one to the left and one to the right, joined at the top, with a horizontal bar in the middle. The problem seems to be how to connect the horizontal bar with the oblique bars in the most effective way.

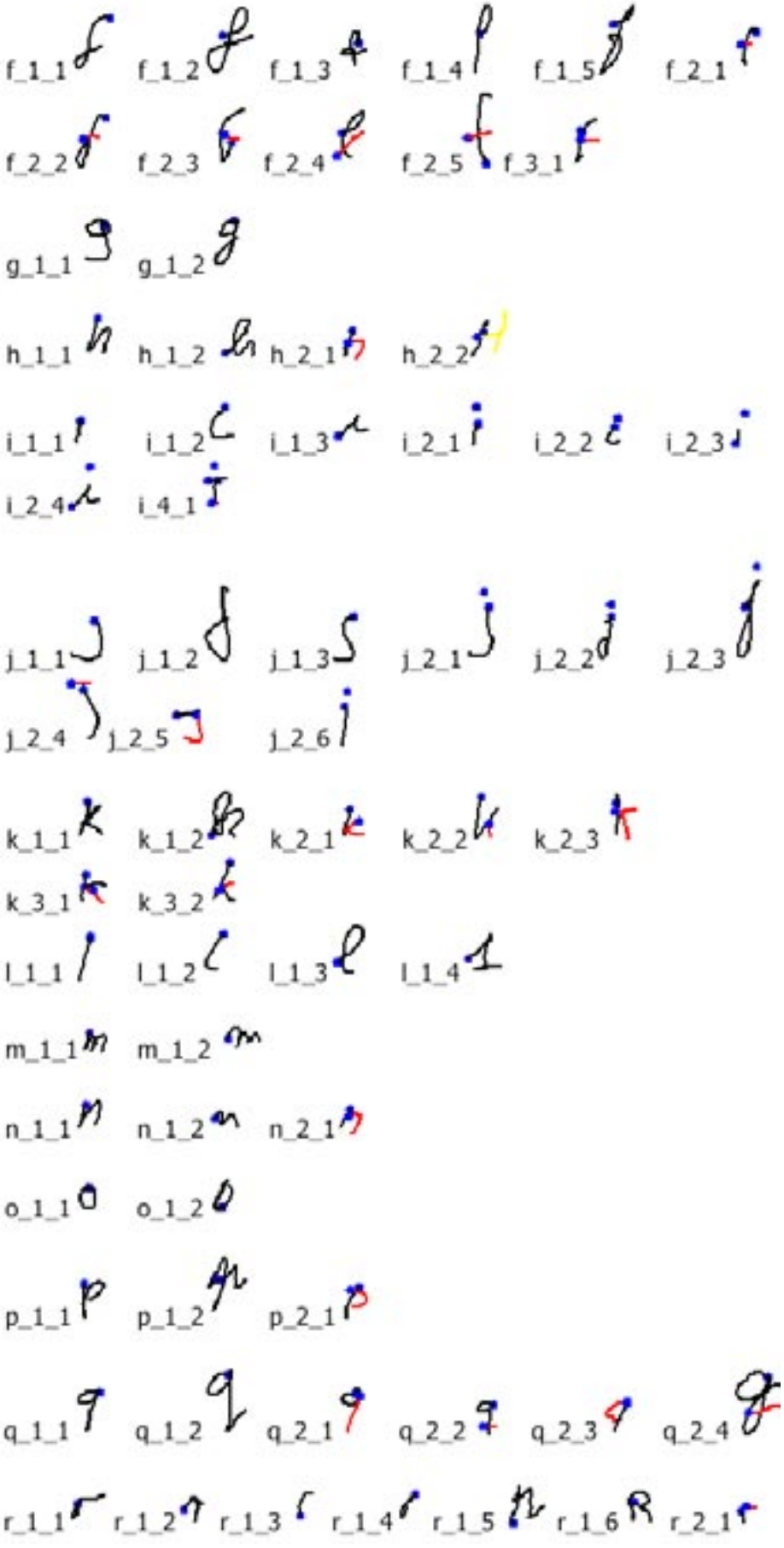
The first four of these are drawn in one arc, the next two in two arcs and the two last in three arcs. The first allograph, called A_1_1, is begun at the lower end of the left-hand bar (as indicated by a dot). The horizontal bar is connected to the lower end of the right-hand bar. A_1_2, on the other hand, is begun at the right end of the horizontal bar and ends at the lower end of the right-hand bar. The third prototype, A_1_3, is an oddity that looks more like a lower case letter. The fourth, A_1_4, is begun at the top. The left-hand bar is drawn first; after reaching the lower end the pen then turns back to the top, to draw the right-hand bar; the horizontal bar is drawn last.

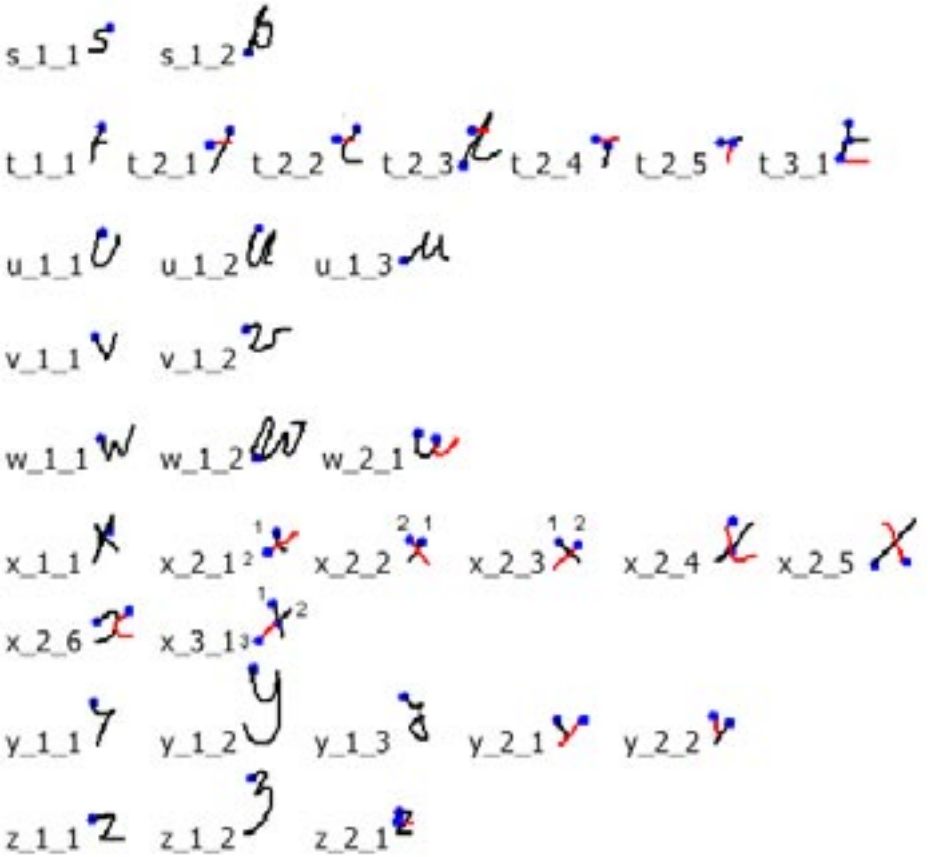
The two-arc allographs differ from the one-arc allographs in that the horizontal bar is drawn separately from the two oblique bars. A_2_1 is begun in the lower left corner and A_2_2 is begun at the top. In the three-arc allographs, all three bars are drawn separately, and in the same order. The difference is that in A_3_1 the right-hand bar is drawn straight downwards, while in A_3_2 it is drawn first to the right and then turns downwards.

Table 1. Upper and lower case allograph prototypes. Each prototype has the code *g_n_i*, where *g* indicates the grapheme, *n* the number of arcs and *i* distinguishes between prototypes with the same number of arcs. Numbers by the graph indicate the order of arcs. Please note that the graphs aren't in any way 'prototypical shapes' but just examples of how the allograph might look




N_1_1 N_1_2 N_1_3 N_2_1 N_3_1
P_1_1 P_1_2 P_2_1
Q_1_1 Q_2_1 Q_2_2
R_1_1 R_1_2 R_2_1 R_3_1
T_1_1 T_2_1 T_2_2
U_1_1 U_1_2
W_1_1 W_2_1
X_1_1 X_2_1 X_2_2 X_2_3 X_2_4
Y_1_1 Y_2_1 Y_2_2 Y_2_3 Y_2_4 Y_2_5
Y_3_1 Y_3_2
Z_1_1 Z_2_1
a_1_1 a_1_2 a_1_3 a_2_1 a_2_2 a_2_3
b_1_1 b_1_2 b_1_3 b_2_1
c_1_1 c_1_2
d_1_1 d_1_2 d_1_3 d_2_1 d_2_2
e_1_1 e_1_2 e_2_1 e_2_1 e_3_1


f_1_1
f_1_2
f_1_3
f_1_4
f_1_5
f_2_1
f_2_2
f_2_3
f_2_4
f_2_5
f_3_1
g_1_1
g_1_2
h_1_1
h_1_2
h_2_1
h_2_2
i_1_1
i_1_2
i_1_3
i_2_1
i_2_2
i_2_3
i_2_4
i_4_1
j_1_1
j_1_2
j_1_3
j_2_1
j_2_2
j_2_3
j_2_4
j_2_5
j_2_6
k_1_1
k_1_2
k_2_1
k_2_2
k_2_3
k_3_1
k_3_2
l_1_1
l_1_2
l_1_3
l_1_4
m_1_1
m_1_2
n_1_1
n_1_2
n_2_1
o_1_1
o_1_2
p_1_1
p_1_2
p_2_1
q_1_1
q_1_2
q_2_1
q_2_2
q_2_3
q_2_4
r_1_1
r_1_2
r_1_3
r_1_4
r_1_5
r_1_6
r_2_1



5.6. Looking for patterns

All TP's but one (DE_04) exhibit allographic variety, that is to say that they use more than one allograph for at least one grapheme. All graphemes exhibit allographic variety except <C, L, O, S, V>; all lower case graphemes have at least two allographs. Several of these are 'curious' and appear only once. But not far into process it was noticed that the most common allographs of some graphemes share certain characteristics. There are three distinct groups among the upper case graphemes and four among the lower case ones. We will concentrate on these groups. Please note that the division into groups only is applicable to some graphemes.

5.6.1. Upper case allograph groups

The allographs that form the upper case groups are united by having a vertical line to the left, except for <A> were it is oblique. It is the way this line is drawn, and how it is connected to the rest of the grapheme, that unites the group members.

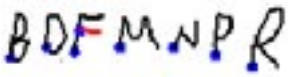
Upper-1

This group is drawn in two arcs. The first arc is begun at the upper left corner of the grapheme and drawn downwards. The pen is then lifted and the second arc is drawn.

Upper-2

Apart from <A> and <F>, this group is drawn in one arc only. The left-hand line is drawn twice, from the upper left corner downwards, then up again.

Upper-3



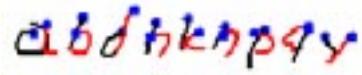
These allographs are all begun in the lower left corner and drawn in one arc except for <F>.

5.6.2. Lower case allograph groups

Lower-1

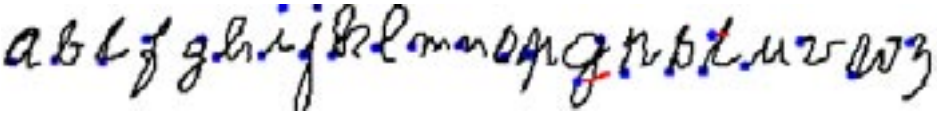
These are the most common allographs among the lower case graphemes. They are the versions normally associated with block letters. They are all drawn in one arc, except for <f> with its horizontal line.

Lower-2



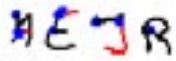
This is another kind of block letters, drawn in two arcs.

Lower-3



These are allographs that are associated with cursive script, all produced in one arc and with 'connectors' protruding from the ends.

Lower-4



Since the test persons were clearly told whether to write in lower or upper case, this group is something of an oddity: it consists of upper case letters written in the place of lower ones. But the phenomenon is too common to be disregarded. As seen in Table 1, there is also the reverse case, where lower case letters are written instead of upper case ones (such as A_1_3), but it is much rarer.

6. Results and analysis

6.1. Allograph group adherence

Do people adhere to the allograph groups? To some extent, yes. _here are a few TP's who use one group only. However, if a clear majority of the graphemes in question comes from one group, the TP is categorized as a user of that group. As to the upper case groups, some TP's use one group only, others mix it with one or two other groups. The strongest link is between <M> and <N>: for these graphemes, a TP never uses allographs from two different groups. Of the 36 TP's, there were 18 classified as Upper-1 users, 13 Upper-2 users, and one single Upper-3 user. The remaining four TP's use an equal number of allographs from groups Upper-1 and 2.

For the lower case groups, the situation is more complex. There are no exclusive users

of a group. A majority of the TP's use Lower-1, but always with elements from Lower-2 and 3. There are 10 Lower-2 users, 6 Lower-3 users and one lower-4 user.

Table 2. Test persons with personal data and allograph group adherence. Presence of other groups is stated in brackets

TP	Gender	Age	Ed. level	Upper case groups	Lower case groups	Number of allographs N^A
DE_03	M	1	4	1	1 (2)	54
DE_04	M	3	5	1	2	47
DE_05	F	4	3	1	3	55
DE_06	M	2	3	1 (2, 3)	3	52
DE_09	M	2	4	1 (3)	1-2	51
DE_10	M	3	4	2 (1)	1-2	52
ES_02	F	2	1	2	1-2	57
ES_03	M	1	4	2 (1)	1-2	54
ES_04	F	1	5	2 (1)	1-2	53
ES_05	M	5	4	1 (2, 3)	1-2	54
ES_06	M	3	3	2 (1)	2	49
ES_08	M	3	4	1 (2, 3)	2 (3)	51
FI_01	F	2	4	1 (3)	1	50
FI_02	F	2	5	2 (1)	1	51
FI_03	F	5	4	1	2 (1)	57
FI_04	M	6	4	1	3	54
FI_05	M	1	4	3 (1)	3	54
FI_07	M	1	3	1 (2, 3)	1 (2, 3)	60
FR_01	M	4	4	2 (1)	1 (2)	52
FR_03	M	3	3	2 (1)	2 (3)	64
FR_04	M	2	3	2 (1)	4 (1)	58
FR_05	M	2	3	1 (1)	3	61
FR_06	F	1	3	2 (1)	1 (3)	59
FR_10	M	5	3	1 (2, 3)	2 (3)	59
GB_01	M	5	(no data)	2	1	49
GB_02	M	4	"	2 (1)	2 (1)	48
GB_03	F	4	"	1 (2, 3)	3 (1, 2)	50
GB_04	F	1	"	1 (2, 3)	1	48
GB_05	M	1	"	1-2	1 (2)	51
GB_07	F	2	"	1-2	3 (1, 2)	51
SE_01	F	2	3	1-2	1 (2)	51
SE_02	M	3	3	1-2	2 (1)	52
SE_03	F	3	4	2 (1)	1	52
SE_04	M	2	4	1 (2, 3)	1 (2)	55
SE_05	M	2	4	1	2 (1)	50
SE_06	M	2	4	1 (2)	2 (1)	50

Age groups	Members
1 15-20	8
2 21-25	12
3 26-30	7
4 31-35	4
5 36-40	4
6 41-45	1

It would of course be a good thing to find matching between the upper and lower case groups, for example between the two-arc Upper-1 and Lower-2 groups. However, there doesn't seem to be any such patterns.

6.2. Geographical distribution

It becomes more interesting when we look upon the geographical distribution of the groups. We find that Upper-1 is the most common group in Germany (5 out of 6 TP's) and Finland (4/6). In France and Spain, 4 out of 6 TP's use Upper-2. The British and Swedish TP's mix groups 1 and 2. Upper-3 is found in all countries.

The distribution of Lower-1 and 2 gives rise to no particular observations. However, Lower-3 is more interesting. It is most common in France, and found in all other countries except Sweden. France is also the only country in which Lower-4 is found. I will try to explain these particularities further on.

6.3. Other factors

There is no particular distribution of allograph groups over gender, age or educational groups. Because of the inexactness of the scale of educational level and the lack of figures from the UK, we will disregard this factor entirely.

6.3.1. Number of allographs as a measure for neatness

However, an interesting observation might be made about the number of allographs N^A that the test persons use. As we remember, there are 21 upper case and 26 lower case graphemes that exhibit allographic variety. For these 47 graphemes, a TP may use anything between 47 (DE_04) and 64 (FR_03) allographs. Could N^A be seen as a measure of the 'neatness' of the handwriting, or if you will, the extent to which the writer has formed her own personal alphabet? We will see how N^A is distributed over the different groups (Table 2). It is of course advisable not to rely too much on such a small and unevenly distributed amount of data, with only six TP's from each country and an uneven number of members of sexual and age groups. But some trends might begin to show.

Table 3. Average N^A of different groups

National average N^A		**Average N^A of age groups**		**Average gender N^A**	
France	58,8	15-20	54,1	Men	53,4
Spain	54,3	21-25	53,1	Women	52,8
Finland	54,3	26-30	52,4		
Germany	51,8	31-35	51,3		
Sweden	51,7	36-40	51,0		
UK	49,5	41-45	54,0		

As we can see, there is a considerable range in the national average N^A. The by far highest is found in France, which could be due to the presence of all four lower-case groups, but also that there are a lot of 'odd' allographs. The neatest handwriters are found in Germanic countries. There is a trend of falling N^A as the writers get older, even if the sole member of age group 6 doesn't fit in. It seems as though younger persons are more careless writers; or to put it more nicely, they haven't yet established a personal alphabet. Men use a slightly higher number of allographs than women.

7. Discussion

7.1. The Latin and Germanic style groups

Could these trends be unified to say something about the entire European situation? Again, it is of course advisable not to count too much on such a small material. Nevertheless, I would like to propose that in these results there is a visible divide between the Latin and Germanic spheres. The Latin group, consisting of France and Spain, is characterized by its use of the Upper-2 group. Within the group, France is separated from Spain by the higher presence of the cursive Lower-3 group.

The Germanic group is divided into two subgroups. There is Finland and Germany, with a high presence of Upper-1; Finland is set apart by its higher number of allographs. Sweden and UK mix groups more freely, but still have a lower number of allographs. It seems that Swedish and British writers stick to the allographs they have chosen, even if they come from different groups.

7.2. Two extremes: France and Sweden

Despite the fact that the writers were told to write block letters, there is a presence of letter forms belonging to cursive script (Lower-3) in the samples. It could be that these writers never have been taught the block letter forms that we take for granted, and that for them cursive script is the natural form of expression.

In the Germanic group, there is a presence of cursive script everywhere except in the case of Sweden. It seems that Swedes either have the ability to switch easily between cursive and block letters, or that the teaching of cursive script has been abandoned in schools, or both. In France, on the other hand, some writers may even be unable to write in block letters, such as FR_04, who reverts to using upper case forms instead of lower case. To sum it up, it seems as if cursive script is more important in the countries on the European Continent, and even Finland and the United Kingdom, than one might have suspected.

8. Conclusion

The work undertaken here is in itself only a pilot study for the possible project of categorizing the entire Decuma database of handwriting samples, since only about one-fourth of the European and none of the American material was studied. This would give further evidence as to the reliability of the trends shown in this paper. Since the manual allograph sorting process used here turned out to be time consuming, some kind of automatization would be of use.

It is of course of great interest to see to which extent these findings might aid the development of HWR systems. It is clear that you always have to be open for the possibility that writers switch between block letters and cursive, the special conditions in each country, and to the higher sloppiness of younger writers. One might imagine an HWR-system that could be set according to the user’s nationality, to limit the number of allographs that need to be considered. If reliable allograph-based style categories are found, they may be combined with other kinds of categorization to give each handwriter a 'style profile', after which the recognizer could be adapted. This might also be used for purposes of identification and safety.

Solving the problems of recognizing cursive script (which is perhaps necessary when having countries such as France, where cursive holds such a strong position, in mind), remains the great challenge within the field of HWR. The methods described in this paper may eventually be applied to a database of cursive script. It is recommendable that the collection of data should be as unconstrained as possible, to give an image of the way handwriting may be done in real life. However, you still have to be open to the infinite variety of handwriting.

9. References

Allén, S. 1971. Introduktion i grafonomi: det lingvistiska skriftstudiet. Stockholm: Almqvist & Wiksell

Bote-Lorenzo, M. L., Dimitriadis, Y. A. & Gómez-Sánchez, E. (2001). Allograph extraction of isolated handwritten characters. Proceedings of the 10th Biennial Conference of the International Graphonomics Society. Nijmegen: IGS

Crettez, J. P. 1995. A set of handwriting families: style recognition. Proceedings of the 3rd International Conference on Document Analysis and Recognition, 489-494. Montreal: Computer Society Press

Davis, T. 1994. The acquisition of handwriting in the UK. http://www.birmingham.ac.uk/english/bibliography/handwriting/new_web_pages/acquisition.htm/

Gelb, I. J. 1963. A study of writing. Chicago: The University of Chicago Press.

Plamondon, R. & Srihari, S. N. 2000. On-line and off-line handwriting recognition: a comprehensive survey. IEEE Transactions on Pattern Analysis & Machine Intelligence, vol 22 no 1, 63-84

Sampson, G. 1985. Writing systems: a linguistic introduction. Stanford: Stanford University Press

Schomaker, L. 1998. From handwriting analysis to pen-computer applications. Electronics & Communication Engineering Journal, June 1998, 93-102

Schomaker, L., Abbink, L. & Selen, S. 1994. Writer and writing-style classification in the recognition of on-line handwriting. Proceedings of the European Workshop on Handwriting Analysis & Pattern Recognition. London: The Institution of Electrical Engineers

Vuurpijl, L. & Schomaker, L. 1997. Finding structure in diversity: a hierarchical clustering method for the categorization of allographs in handwriting. Proceedings of the 4th International Conference on Document Analysis and Recognition, 387-393. Piscataway, NJ: IEEE Computer Society

10. Swedish equivalents of some English terms

affine transformation affin transformering
allograph allograf
arc 'båge', streck
block letters tryckbokstäver
cursive skrivstil, kursiv stil
ductus duktus
grapheme grafem
graph graf
hand, handwriting style handstil
handwriting recognition handskriftsigenkänning
idiograph idiograf
stroke penndrag
variability föränderlighet
variation variation
variety mångfald
writer skribent