

# On The Structure of Style Space for Documents

**Rhys Price Jones**

Rochester Institute of Technology  
email: rpjavp@rit.edu

**J. Fernando Naveda**

Rochester Institute of Technology  
email: jfnics@rit.edu

**Paul Roetling**

Roetling Consulting  
email: illusion@rochester.rr.com

**Steven J. Harrington**

Xerox Corporation  
email: sharrington@crt.xerox.com

**Nishant Thakkar**

Rochester Institute of Technology  
email: knetsar@charter.net

## Abstract

We identify three aspects of style pertaining to documents. The first of these we call literary style and it includes the word and sentence constructions and choice of illustrations traditionally associated with authorship. The second we call informative style and it includes formatting and iconic choices that convey additional information such as the document's genre or corporate identity. The third aspect of style covers the degrees of freedom remaining for the author and is used to convey the author's intent. Literary style is the realm of academic scholarship and discourse and is beyond the scope of the present article. But corporate and intent style can be quantified by measuring many different attributes. For example, density of text, colorfulness of images, regularity of positioning of images, diversity of font and typeface, all contribute to the document's overall style. Indeed, we have identified more than 150 different value functions, each of which can be measured, and each of which can contribute to a document's overall stylistic appearance. Measurement of these value functions effectively places a document as a point in a style space. But the 150 value functions are not independent. A heuristic approach is described for investigating the possibility of finding basis vectors for intent space.

## Introduction

Jonathan Swift, in his Letter to a Young Clergyman on the 9th of January 1720, asserts:

Proper words in proper places, make the true definition of style.

In other words, Swift is asserting that style deserves no more attention than the selection of words and the context within which those words are placed. To us, that constitutes **literary** style, and distinguishes great wordsmiths from ordinary mortals. Oscar Wilde, in *The Importance of Being Earnest*, has a differing opinion:

In matters of grave importance, style, not sincerity, is the vital thing.

Wilde recognized that there is more to style than just the selection of effective words and word arrangements for the more effective communication of ideas. Many in this age of multifarious multimedia, would agree with Wilde. Has

superficial style overtaken substance in the realm of communication?

It is not our intention in this paper to contribute to that particular debate. We will not presume upon the realm of literary criticism and will have nothing to say about literary style. Instead we will consider the space spanned by all aspects of style except the sincerity. Some of that style may be mandated by the owners of the medium. For example, as we prepare this very document for publication by AAAI, we must abide by a number of dictates. The rules ensure a consistent appearance for all documents submitted to this workshop. Nevertheless, a great deal of flexibility remains at the disposal of the author and designer and can be used in various ways to further their intentions. We will describe and investigate this intent space.

## Literary vs. Format Style

In preparing a document for presentation, the author makes a number of decisions. Literary style comprises selecting the concepts to be conveyed, the language in which to express them and the choice of words and sentence structures. Beyond those choices there are choices of font, spacing, positioning, color, margins and other parameters that specify the format and layout of the document. These format parameters define a style space and the specific choices made will associate the document with a point in the format style space.

What distinguishes the style of Dickens from the style of Dumas? Not only is English significantly different from French, but the authors also use different ways of arranging different combinations of words to convey different nuances. Style is more than the concepts conveyed, it includes the exercise of all the options in how to convey them. The essence of each **literary** style lies entirely in the word selection and arrangement. Literary style is essentially independent of medium. The literary portion of Dickens on paper is the same as the literary portion of Dickens on the Web is the same as the literary portion of Dickens on audiotape. It might even be argued that an adaptation of Dickens for cinema or television or, God forbid, a video game would also preserve the literary style.

Format style is, on the other hand, media dependent. The corporate look and feel of a corporation on paper may differ considerably from its look and feel on the web. One of the

consequences of the work we discuss here and elsewhere is that it may be possible to automatically convert documents from one medium to another and preserve the style.

There is a small overlap between format style and literary style. Not only do corporate guidelines impinge on decorative aspects of a work; they can influence some literary choices too. Contributors to the New York Times must use Standard American Spelling and conform to other house rules concerning inclusive language and so forth. Nevertheless, we believe the format aspect of style space is well defined and clear.

### **Value Properties**

Designers and authors of documents make many decisions, quite divorced from literary style considerations, about how they will present their information. The overall appearance resulting from those decisions contributes to the essential format style of the document. As noted above, we make no attempt to quantify the literary aspects of style, such as the author's choice of words. We will, however, endeavor to quantify the remaining aspects of the concept of style. We look for features that can be measured. We call these value properties because they are associated with the notion of goodness. Some examples are distinguishability, legibility, economy, balance, group identity, uniformity and how eye-catching something is. There are lots of possible value properties; every rule in a document design book is a potential property. There is typically a hierarchy of properties as one moves from the specific to the general. Some properties can be directly measured; others can only take on a value when considered relative to other features on the page. For example, one may directly measure the size of a font used in a certain passage; but to elicit a value for how it contributes to the eye-catching measure requires knowledge of font sizes in surrounding text, as well as other properties such as color, face, and viewing distance. So some value properties can be empirically measured, whereas others are computed in terms of other value properties. The definition of a set of value properties that contribute to a measure of document aesthetics are described in (Harrington *et al.* 2004). Value properties are calculated on the format style parameters, and therefore provide a mapping from the high dimension format style space to a lower dimensioned space of goodness values. However, because of correlations among the value properties, we do not expect them to provide a basis for this space. We expect a much smaller set of effects or goals that the author tries to achieve.

### **Informative Style**

There may be constraints on authors that require them to conform to guidelines mandated by their publishers or the owners of the media on which their works will appear. For example, the overall appearance of articles in the New York Times must be consistent with the Company's policy. It would not be permitted for an author to intersperse varieties of typefaces and fonts without a good reason. Some of these rules are designed to support certain value properties (e.g. to ensure legibility), but rules can also be used to achieve

a corporate identity. Style choices can therefore convey information (such as which company created the document). Communications from the Xerox Corporation are expected to conform to company policy, use appropriate letterhead, etc. Thus the author and designer do not have completely free rein with the non-literary parameters of style. In a similar vein there are rules that inform us of a document's genre. A letter, a memo, a newspaper article, a technical report, and a resume all look different. That difference gives us information that we use to establish context and expectation for document.

### **Intent Space**

We have previously reported (Harrington, Naveda, & Price Jones 2001) that what really guides the choices of a web author is an assignment of importance to certain value properties over others so that the document communicates more effectively. This assignment of importance is a reflection of the author's intent. For example, in designing a paper phone book, the intent might be to economize on resources. This would lead to an adoption of values for value properties that would lead to a very spartan overall style. Font sizes would be small, spacing would be very cramped, and the density of text would be unpleasantly high. For another example, consider the design of a web page to cover a newspaper story. This time, the intent is to convey the gist of the story in a very prominent short headline, and to provide the details in a less exaggerated component. The designer might also be concerned about download time and this might be reflected in the presence or level of detail of any illustrations.

We define the specification of the relative importance of the values as intent. If we assume that there are a small number of high-level values or goals to be achieved, then there are a corresponding small number of intent weights. The space spanned by the small number of high-level value functions we will call intent space.

Some aspects of the overall location of a document in style space are thus attributable to author intent (Harrington 2003). The overall style of a document is, at least in part, a consequence of assignments to value properties, corresponding to a point in style space.

### **An Experiment**

We collected a set of several hundred diverse web documents. We chose to use web documents because of availability and variety. The web pages included software documentation, notifications, news reports, advertisements, biographies, corporate home pages, transcripts, product listings, product descriptions, virtual tours, resumes, specifications, contact lists, forms, reports, search results, menus, glossaries, tutorials, recipes and personal web pages. We subjected each to our automated measurement of 150 value functions and noted these values. Some examples of the value functions are listed in Table 1.

Next we assembled a group of twenty graduate students in the RIT Computer Graphics Design program to cluster these documents according to similarity of "style". It is important to note that we made no attempt to define to our volunteers

Average Character Size	Density
Average Relative Line Separation	Document Left Alignment
Average Spatial Separation	Document Table Density
Bullet Density	Document Top Alignment
Change in Interest	Document Verbosity
Color Harmony	Foreground Fraction
Colorfulness	Group Alignment
Consistency of Continuity	Interest
Consistency of Luminance	Intimidation
Consistency of Size	Italic Density
Consistency of Structure	Limitation of Color
Consistency of Visual Importance	Number to Word Ratio
Content Neatness	Page Golden Aspect
Content Novelty	Reading Level
Cost	Spatial Coherence
Document Consistency of Style	Text-Image Balance
Document Continuity	Text Line Retrace
Document Grouping Number	Visual Left-Right Balance
Document Horizontal Regularity	Visual Vertical Balance
Document Information	Visual Uniformity

Table 1: Example value functions

what we mean by style. We just asked them to group together, in any quantities they pleased, documents that they thought had a similar style to each other.

Finally, the cluster information was used in an attempt to perform a principal components analysis of the clustered points in intent space. Further work is warranted, but preliminary results support our intuition that intent space may perhaps be of fairly low dimension.

Volunteers suggested “convenience”, “eyecatching”, “comfort”, and other similar terms as criteria for their clustering. We suspect that there may be a fairly small set of vectors in intent space that constitute a basis. These might be thought of as fairly sweeping and inclusive value functions. Concepts as elusive as “aesthetics” (Harrington *et al.* 2004) or “ease-of-use” may well serve as basis vectors for intent space.

Our experiment was conducted on a set of diverse web documents. We can imagine conducting a similar experiment on a set of diverse printed documents, or on a set of diverse PowerPoint presentations. Indeed, it is possible to perform this same analysis for sets of diverse documents in any medium. Further experiments are needed, but at this stage we feel it is reasonable to conjecture that similar results will be obtained for each medium: style will be measurable in terms of large numbers of value properties; but the entirety of the space spanned by all possible assignments to those value properties will have a basis consisting of a small set of intents.

### Intent Basis

A true-factor analysis of the value property measurements was applied to our set of web documents. Typical results are shown in Figure 1. This was due to value function definitions being somewhat arbitrary, allowing the building in of correlations that the analysis was sensitive to. Also, the basis vectors derived in this way did not map directly into intuitive

values that would be useful for human interaction. While it was possible to define an intuitive basis and transform between it and the true factors, we found that for the purpose of document design it was simpler to just work with the intuitive set directly. The value properties have been applied to the problem of automatic document layout (Purvis 2002; Purvis *et al.* 2003). The high-level value properties to which we applied intent weighting were:

1. Aesthetics, the beauty of the appearance
2. Eye-catching, the ability to grab attention
3. Interest, the ability to maintain attention
4. Communicability, the effectiveness in conveying information
5. Comfort, which includes familiarity and lack of irritation
6. Convenience, how well the document supports its use
7. Economy, how low the cost of production and use in time and money

These rather abstract concepts are quantified in terms of simpler measurable properties. For example, aesthetics is defined in terms of features noted as good design and the lack of features considered bad design. These include alignment, balance, the white-space fraction, uniformity and so on (Harrington *et al.* 2004). The definitions for the value properties are primarily motivated by guidelines for good document design (Turnbull & Baird 1964; Topping 1990; Collier 1991).

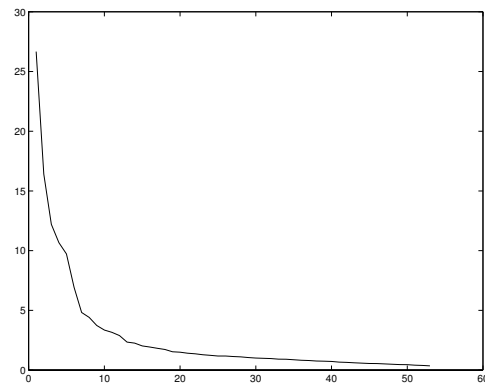


Figure 1: True Factor Analysis

Note that this basis is not orthogonal. There is substantial overlap between the concepts of eye-catching and interest. These properties share some simpler property measures in their definitions (although with different weights). There is also a great deal of overlap between the ideas of comfort and convenience. In fact one might wish to combine these into a single ease-of-use measure.

## Changing Media

If it turns out that a single set of intents can prevail as basis vectors across multiple different media there are important consequences for the field of automatic document conversion. Biber (Biber 1995) has investigated a concept he calls **register** (roughly corresponding with our notion of intent), in particular how variations in register manifest themselves between different languages. We, however, confine ourselves to a single language but consider manifestations of intent in different media.

When we move from the space of web documents to documents in other media, it is possible to express those same intents by assignments to different value properties appropriate for the new media. The resulting document may look very different from the original, but it carries the same intent by the authors and designers. In other words, a style for one particular medium corresponds to another style for another medium. But, crucially, the two styles are a manifestation and indication of the same intents.

We believe this view of style can be applied to areas besides document formats. In the world of art, one might paint a picture to convey a message (for example a scene from a religious story). In addition to the choices of what elements go into the scene, there are choices of style that indicate how those elements are presented. Some choices correspond to the school or genre (e.g. impressionistic) and to the identification of the artist. Choices also govern how the work affects the viewer (does it catch the eye, is it comfortable to view, does it evoke emotion). One might also imagine the same message conveyed in a different media (such as sculpture). In this case the same intent might be expressed in different style decisions (perhaps using size or texture instead of color)

A document prepared for one type of media can be analyzed and positioned accurately within the style space of that medium. We locate that document in intent space via a basis of intents for that medium and a transformation matrix for those intents to value functions in that medium. Armed with the intent information, we then proceed to another medium and via the transformation matrix for that new medium we convert the same intents to a new set of value function readings for the new medium. By adjusting those values for the document, we format it so that it is rendered in the new medium with the same intents, and thus the same value style, as the original document on the original medium. The new document will be very different from the original, but will have the same impact on the viewer as the original document's author and designer intended for the original on its medium.

An important application is the cell-phone and PDA delivery of web pages. A highly adorned web page will not transfer to the PDA without replacing its adornments with something that will provide a similar level of decoration when rendered on the PDA, but the original intent still prevails. Preparation for a cell-phone is a radical change but as an increasing number of them have screens, the meaning of intent might change from a purely visual concept to something that includes voice and sound. So for example, the rendition of the "communicability" or the "eye-catching" value functions

will certainly change when delivered to a cell-phone.

It will be possible to automatically convert web documents to printed documents that will have the same impact as the authors intended. It will be possible to automatically convert any kind of document for display on a small handheld screen such as might be found on a cellular telephone; and the resulting display will be in a style that corresponds naturally to the intentions of the original authors. It will be possible to convert a printed document, such as this, to a PowerPoint presentation designed to have the same impact as the original paper!

## Conclusion

The science of intents promises a new approach to automated media conversion for informative documents. If correspondences can be identified between author intents and measurable properties of documents on individual media, then those correspondences can be used to convert documents to other media and produce equivalent documents on different media that conform to the original's intent.

## References

- Biber, D. 1995. *Dimensions of register variation*. Cambridge University Press.
- Collier, D. 1991. *Collier's Rules for Desktop Design and Typography*. Addison-Wesley.
- Harrington, S.; Naveda, J.; Price Jones, R.; Roetling, P.; and Thakkar, N. 2004. Aesthetic measures for automated document layout. In *the ACM Symposium on Document Engineering*.
- Harrington, S.; Naveda, J. F.; and Price Jones, R. 2001. Towards a science of document intent. In *ICDAR 1st International Workshop on Web Document Analysis*.
- Harrington, S. 2003. The application of intent to style. In *Doing It With Style Workshop*, 55–59. Acapulco, Mexico: IJCAI.
- Purvis, L.; Harrington, S.; O'Sullivan, B.; and Freuder, E. 2003. Creating personalized documents: An optimization approach. In *the ACM Symposium on Document Engineering*.
- Purvis, L. 2002. A genetic algorithm approach to automated custom document assembly. In *the 2nd International Conference on Intelligent Systems Design and Applications*. Atlanta, GA: ISDA.
- Topping, S. M. 1990. *Graphic Design and Color in Today's Office*. Kodak Publication W-628.
- Turnbull, A. T., and Baird, R. N. 1964. *The Graphics of Communication*. New York: Holt, Rinehart & Winston.