Revealing semantics using subtle typography and punctuation

Kumaran Sathasivam, S.K. Venkatesan and Yakov Chandy

Abstract

The semantics of a language has deeply nested structures, which is revealed by typography using a hierarchy of paragraphs, with different font sizes and styles. At the paragraph level, a paucity of typographical features forces us to use punctuation heavily to reveal semantics. Paragraphs are further broken into smaller semantic units such as sentences using end-punctuation and initial capitals. Sentences are further broken down using semi-colons, colons, commas and hyphens into even smaller chunks. Word-spaces are used to break language into the smallest atoms of semantics, namely words or phrases. In this paper, we look at newer devices, both typographic settings and punctuation elements, that can disambiguate and reveal deeply nested semantic structures.

> "To reveal art and conceal the artist is art's aim." — Oscar Wilde

1 Introduction

The indivisible elements, the atoms as it were, of the written forms of languages like English are letters. But reading text built of these basic units alone is a difficult exercise. An elaborate set of conventions, auxiliary symbols (punctuation symbols), spacing and typography are used in publishing today as aids to reading, as removers of obstacles to understanding. The salutary effects that these jointly have will be readily appreciated by a comparison of the first two sentences of this paragraph with a version written entirely in uppercase and unencumbered by punctuation symbols or word spaces (Figure 1). Text was, in fact, written historically thus. There was no spacing, paragraphing or punctuation in manuscripts before the development of printing (Boorstin, 1983).

Uppercase characters are still used amidst lowercase characters. Uppercase letters are used to display title and headings in a prominent way, similar to how they were used to display text in ancient Roman buildings. Uppercase letters are also used as the beginning character of prominent names (proper nouns) and later evolved to shorter form initials, abbreviations and acronyms. They are also used at the beginning of a sentence as a device to prominently mark the beginning. Special uppercase characters are used as dropped capitals in some books as a decorative element at the beginning of a chapter or other part of the document.

THEINDIVISIBLEELEMENTSTHE ATOMSASITWEREOFWRITTEN ENGLISHARELETTERSBUTREADING TEXTBUILTOFTHESEBASICUNITS ALONEISADIFFICULTEXCERCISE

Figure 1: Text in uppercase only, stripped of punctuation symbols and word spaces.

The lowercase letters started initially in the cursive italic form, but now they have evolved their own upright roman form. Bold and display fonts have more or less replaced the necessity of displaying title and heading in uppercase characters. Interestingly, another form known as the small-caps has evolved from capital letters with its own variety of lowercase characters, an interesting hybrid in typography.

Fonts may be broadly classified into serifs and sans-serifs. Slab serifs are a significant subgroup of serif fonts, usually with relatively thick strokes and with flat, rectangular serif shapes. Apart from these we have other categories of fonts used, such as monospaced fonts, used to display telegrams (such as in Wikileaks messages) or computer code and other types of fonts such as Gothic and script fonts to reveal specific content or context.

At the paragraph level, the document title, headings, paragraphs, footnotes are distinguished by typographic elements such as use of different font styles and font sizes. However, once we come down to the paragraph level, except for occasional embellishments with bold and italic, upper- and lowercase characters, we are left with only the punctuation marks to describe the underlying structure of the sentences. The role of punctuation transcends that of merely providing assistance for the eye; it now encompasses interpretation of text. In other words, punctuation is vital for the meaning of written material.

It is therefore not surprising that the first known use of punctuation is related to a system that was used to help the delivery of speeches from written texts. This system dates back to the fifth century BC, when the Greeks introduced vertically arranged dots in text. Subsequently, when Greek playwrights (for example, Aristophanes and Euripides) wrote drama, they used symbols to distinguish the ends of phrases so that the play's cast knew when to pause. Even relatively recently, school children have learned to associate punctuation marks with pauses in reading. One mnemonic poem actually quantifies the duration of the pause associated with each major punctuation symbol:

Figure 2: The dramatic change in meaning that a single comma can introduce is illustrated by this example. [Source: `http://www.themodernausten.com/2012/09/04/teacher-tuesdays-9-4-12/`]

Charles the First walked and talked half an hour after his head was cut off.
Charles the First walked and talked; half an hour after, his head was cut off.

Figure 3: In this example, the meaning of the first sentence is intriguing though rather macabre. A semi-colon and a comma change the meaning of the same string of words entirely, to something more likely. [Source: The American Printer, 1885 edition]

> The stop point out, with truth, the time of pause
> A sentence doth require at ev'ry clause.
> At ev'ry comma, stop while one you count;
> At semicolon, two is the amount;
> A colon doth require the time of three;
> The period four, as learned men agree.

The semantic role of punctuation is easily highlighted using a couple of well-known facetious examples (Figure 2, Figure 3). Thus a particular string of words can have different meanings, depending on the punctuation. In fact, a particular sequence of words can even have two opposite meanings (Figure 4).

This is not paradoxical. If such texts with different meanings are delivered *orally*, they are spoken very distinctly, depending on the intended meaning. The written forms, since they are reduced to identical strings of symbols, require auxiliary support, *marking up*, in the form of punctuation to make the distinction.

2 Evolution of punctuation

Punctuation was developing rapidly at a time when large numbers of copies of the Bible were produced by copyists in Europe, in the fifth century AD. These copies were designed for reading aloud, and so a range of marks were introduced in the text. An early

Woman, without her man, is nothing.
Woman: without her, man is nothing.

Figure 4: The first of these sentences emphasizes the importance of men; the second, the importance of women.

version of initial capitals (the use of lowercase letters to write sentences, except for the first letter of the sentence, which is in uppercase) made its appearance at this time. In the eighth century AD, Irish scribes introduced the practice of separating words. This was a major step in semantics, as now words began to have a standalone existence, making them candidates for a study on their own, reinforcing a socially shared context, through the use of dictionaries. The English language also evolved as an isolating language, making isolation of words possible, but at the same time increasing the importance of the position of the words, creating a need for position-based syntax and grammar. Over the next several centuries, the movement was from words to phrases and several systems of punctuation appeared, some of them disappearing after a spell of popularity, others persisting unchanged or evolving with time.

The use of movable type and the rise of printing in Europe in the 15th century led to an increase in the amount of material printed and in its readership. The printing press spread to hundreds of cities in Europe within decades. It is estimated that by 1500 the printing presses of western Europe had produced 20 million volumes. The need for a standard system of punctuation was keenly felt. Two printers of Venice, both named Aldus Manutius, one the grandson of the other, are credited with the invention of such a system. To the printers Aldus Manutius are attributed the development of punctuation practices that continue to this day, such as the one of ending sentences with full stops, and the development of symbols such as the modern comma. The younger Manutius said in 1566 that the main object of punctuation was the clarification of syntax. The trend of punctuation reflecting sentence structure continued. Notable in this context is Ben Jonson's book English Grammar, published posthumously in 1640, which provided the foundation for the punctuation rules followed today. This is not to say that there is total agreement about the rules of punctuation — there is still some range in punctuation usage.

Printing presses spread further, and in the 16th century they produced between 150 and 200 million copies. In the 19th century, the hand-operated press was replaced with the steam-powered rotary

press, which allowed printing to be performed on an industrial scale.

3 Punctuation rules and style manuals

As early as the late 17th century, manuals (such as Moxon's Mechanick Exercises, 1683–1684) were being produced for the printing trade. With the passage of time, these increasingly addressed the general reader. One of the most successful printing manuals of the 19th century, The American Printer, was published in 18 editions between 1866 and 1893. The preface to the first edition of this manual said that 'Authors and publishers, as well as typographical amateurs, may consult the volume with profit; and indeed, any intelligent person will find it a serviceable companion.' The American Printer (15th edition, 1885) touches very lightly upon the subject of punctuation when it outlines the work of the proof-reader: 'The compositor is bound to "follow the copy," in word and sentiment, unless, indeed, he meets with instances of wrong punctuation or false grammar, (and such instances are not rare,) which his intelligence enables him to amend.'

Just what correct punctuation might be was being defined by academic presses around this time and in the early 20th century, by which time they were drawing up their own rules or standards for typography. Horace Hart, controller of the Oxford University Press (OUP), had worked some three decades at printing establishments, compiling best practices over this period. In 1893 these were printed as a single broadsheet page for use at the OUP. They developed over the years, and were published in 1904. Hart's Rules quickly became a source of authoritative instructions of not just typesetting style but also English usage, grammar and punctuation. Similarly, in the 1890s a proofreader at the University of Chicago Press had drawn up a single sheet of typographic fundamentals. In 1906, the Chicago Manual of Style (CMS) was published as a book. The CMS16 (2010) is now in its 16th edition, and its guidelines have been shaped by ideas from both within the press itself and outside. The print version of CMS16 has more than a thousand pages, and there are more than 2000 hyperlinked paragraphs online. The Web site of the CMS says that it 'has become the authoritative reference work for authors, editors, proofreaders, indexers, copywriters, designers, and publishers'.

Editors and other users of style manuals tend to follow their prescriptions slavishly although the manuals themselves point out that there is nothing hard and fast about their 'rules':

> As always, most Chicago rules are guidelines, not imperatives; where options are offered, the first is normally our preference. Users should break or bend rules that don't fit their needs, as we often do ourselves. Some advice from the first edition (1906), quoted in the twelfth and thirteenth editions and invoked in the fourteenth, bears repeating: "Rules and regulations such as these, in the nature of the case, cannot be endowed with the fixity of rock-ribbed law. They are meant for the average case, and must be applied with a certain degree of elasticity." (CMS15, 2003)

The desire to adhere zealously to the guidelines of style manuals possibly arises in response to the intricacy of the rules: the CMS and the New York Public Library's style manual (Sutcliffe, 1994) devote close to 50 pages each to the use of punctuation symbols alone. It is evident that the prescriptions have been drawn up with great thoroughness to deal with every conceivable situation that may arise when using punctuation symbols.

Perhaps as a caution against overenthusiastic enforcement of the recommendations, CMS15 reminds users that '[p]unctuation should be governed by its function, which is to promote ease of reading. Although punctuation, like word usage, allows for subjectivity, authors and editors should be aware of certain principles lest the subjective element obscure meaning. The guidelines offered in this chapter [Punctuation] draw for the most part from traditional American practice.' In other words, the essence of punctuation is to disambiguate.

4 Semantic inadequacies in current methods of punctuation and typography

The rules of punctuation are evolving continuously. New, revised editions of style manuals are published periodically. To quote words from Wikipedia, 'The rules of punctuation vary with language, location, register and time and are constantly evolving.' This continuous evolution of punctuation is due partly to developments in the language and partly to the fact that the 'rules' laid out in style manuals are often merely descriptive. Further, some aspects of these rules are rather whimsical. As a result, there are inadequacies in the prescriptions regarding punctuation. In Box 1 we present two cases where redundancy is present in the rules of punctuation.

In fact, there are serious limits to how well even the traditional function of punctuation, namely, indicating how text is to be read, can be carried out. As the Wikipedia entry on 'Punctuation' puts it, 'Even today, formal written modern English differs subtly from spoken English because not all emphasis and disambiguation is possible to convey in print, even

Box 1: Redundancy in punctuation rules.

- *Demarcation of sentences.* Sentences are clearly marked off using both end-punctuation (full stops, question marks, exclamation marks) and initial capital letters. It would be more rational for a style manual to prescribe the use of a single sentence-separation punctuation symbol.
- *'Which' versus 'that'.* A distinction is made by style manuals, most American ones, between the relative pronouns 'which' and 'that'. These manuals prescribe the use of 'that' with a restrictive purpose, to narrow a category or identify a particular term being talked about. 'Which', on the other hand, is recommended for nonrestrictive use, not to identify a particular item or narrow a class but to add something about an item that has already been identified. The style manuals redundantly prescribe that 'which', when used nonrestrictively, should always be preceded by a comma, a parenthesis or a dash (CMS15, page 230).

Person 1: (Aggressively) Where did you get that?
Person 2: (Inquisitively) What?

Figure 5: Playwrights must provide stage directions to indicate how lines are to be spoken by actors. [Source: `http://www.tes.com`]

with punctuation.' Nowhere are these limits felt more keenly than in play scripts, where the dramatist must liberally add adverbial stage instructions (Figure 5). Furthermore, we present in Box 2 a couple of instances we found where there may be ambiguity even when the rules of punctuation are followed faithfully.

The development of punctuation is proceeding perhaps faster than ever before. New punctuation symbols have been proposed. Houston (2013) describes the deliberate creation of a symbol to convey a mixture of surprise and doubt. This symbol, known as the interrobang, enjoyed some popularity during the late 1960s and early 1970s. It was lost in the transition of the printing industry from hand-set, hot metal printing to phototypesetting. Houston (2013) also describes how, even more recently, a new symbol, the sarcasm mark, was proposed by a blogger who observed that written sarcasm was regularly misinterpreted as sincerity in online interactions. The need for enhanced disambiguation through new punctuation has been both necessitated and facilitated by

Box 2: Ambiguity in punctuation rules.

- *The comma in an unclear role.* The comma is used in a great many situations. Sometimes, it is not clear what role it is playing. For instance, according to one CMS15 rule, 'A word, abbreviation, phrase, or clause that is in apposition to a noun is set off by commas if it is nonrestrictive — that is, omittable, containing supplementary rather than essential information.' The first example provided for such a use of the comma is 'The committee chair, Gloria Ruffolo, called for a resolution'. Another CMS rule describes the use of the Oxford comma: 'Items in a series are normally separated by commas ... When a conjunction joins the last two elements in a series, a comma — known as the serial or series comma or the Oxford comma — should appear before the conjunction. The first example provided for this rule is 'She took a photograph of her parents, the president, and the vice president.' If, however, the sentence 'She took a photograph of her father, the president, and the vice president' is encountered, how is the reader to interpret it? Did she take a photograph of three persons? Or was her father the president, so that she took a photograph of only two persons, namely her father and the vice president? This type of ambiguity is encountered when the Oxford comma is followed.
- *A relative clause ambiguity.* Sometimes it is difficult to distinguish between interrogative and nominal relative noun clauses. An example is the sentence 'I forgot what he asked for.' One interpretation is 'I know what he asked for. But I forgot to bring it.' In this case, the object of the sentence is a relative noun clause. Another interpretation is 'I do not know any longer what it is he asked for.' Here, the object is an interrogative noun clause.

the development of electronic communication. Smileys, emojis, emoticons and sarcasm tagging have all gained widespread recognition and usage through text messages and emails. It is quite possible that these new punctuation symbols will be absorbed into regular print typography.

5 Extending the LaTeX solution for deep structures

Like XML and HTML, LaTeX provides a mechanism for separation between style and content. The user provides hints and hooks for the typesetting engine and the typesetting engine then provides the rendering, incorporating some hyphenation and justification of paragraphs, and paginating the document. There is now adequate computer power to take LaTeX typesetting to higher levels of sophistication.

We note that the rules of punctuation and typography need to be revisited. We believe that the present setting provides opportunities to advance disambiguation, for both the human reader and the machine reader, through imaginative typography and punctuation.

First we construct some simple solutions to existing disambiguation problems, before we move on to the general problem of structure and semantics.

- *The comma in a dual role.* We could disambiguate the sentence 'She took a photograph of her father, the president, and the vice president.' in the following ways: 'She took a photograph of her father ,the president, and the vice president.' Here we try to convey the information that her father is the president by putting 'the president' within parenthetical comma, which is comma that is shifted after the word-space to indicate the parenthetical nature of it. Likewise, when there is a list with items containing coordinating conjunctions, and there is no Oxford comma, we would compress, expand or letterspace the text automatically. The example we have is the sentence 'We had ice cream, fish and chips and strawberries and cream at the tennis match.' We could disambiguate this by writing it this way: 'We had ice cream, fish · and · chips and strawberries · and · cream at the tennis match.' We hope that the typography of LaTeX will be clever enough to distinguish this nesting indicated here through curly braces and provide subtle typographic effects to indicate the nested structure. When Zapf's hz-program squeezed or stretched individual characters by a few percentage points or letterspaced text (to avoid rivers) in the mid-1990s, it was considered sacrilege by purists. Be that as it may, these features, which Zapf referred to as 'micro-typographic features', have entered LaTeX and other typesetting programs. We believe that these features, when used appropriately, will serve the purpose of disambiguation.
- *Relative clause ambiguity.* Can we disambiguate the following sentence? 'I forgot what he asked for.' As mentioned above, this has two possible interpretations:
  1 I *forgot* that he asked for something;
  2 I cannot remember *what* it was that he asked for.

  The solution:
  1 'I forgot¿ what he asked for.'
  2 'I forgot what¿ he asked for.'

  We hope that the introduction of new punctuation after 'forgot/what' solves the problem. It may require subtle education of both the readers and authors.
- *Sentence break.* The sentence demarcation problem from Box 1. A simple solution would be to introduce a hidden (nonprinting) demarcation symbol between sentences. We could introduce a new character, say a square dot, '▪' (a small version of the Halmos square box). This would provide unambiguous information to a 'machine reader'.

At present, with all the progress in AI, especially Deep Learning, there are now quite a few natural language processing (NLP) libraries that could also be used by the LaTeX engine to interpret the input and produce typeset output as desired, with options to switch on/off such features by the authors of LaTeX. Such interactive NLP systems can also query the author when in doubt. Of course, all this becomes feasible only if we have a rich set of font families with subtle variations and a set of new glyphs for proposing new punctuation marks. A sentence tagged with parts of speech by NLP, with additional indicators introduced by LaTeX authors, can provide a wide scope for LaTeX to typeset the output with subtle typographic variations revealing the underlying structure. Noun phrases and verb phrases can be indicated by either a different font-face variant and/or replacing word-spaces within the phrases by a middle dot, '·', as in the example, 'We had ice cream, fish · and · chips and strawberries · and · cream at the tennis match.'

Zapf's 'micro-typographic features' have entered LaTeX and other typesetting programs, and these features along with newer typefaces for representing grammatical structures and parts of speech will help us evolve language and typography to higher levels.

6 Conclusion

Evolution of style, typography and punctuation have been studied in some detail here. In this article we have also indicated the flavour of the interesting new world that lies before us, while at the same time indicating the need for a huge family of fonts in the great tradition of Knuth's computer modern font or even a myriad of fonts in the MetaPost tradition. We also need a repertoire of new glyphs for punctuation, to express ourselves fully in writing as we would indicate in speech (in tonal variations) and actions (hand, head movements) in our real life. There are quite a few glyphs in musical notation that can give us the inspiration and direction: ♪ 𝅗𝅥 ♫ 𝅝 ♩ ♮ ♭ ♯. Music is an interesting area to start searching for, as it was here in our conception of music that our baby steps in speech and language took shape; there could be something interesting that we left behind in our early creative outpourings.

References

Boorstin, D. J. (1983). *The Discoverers*. Random House.

CMS15 (2003). *The Chicago Manual of Style*. The University of Chicago Press, 15th edition.

CMS16 (2010). *The Chicago Manual of Style*. The University of Chicago Press, 16th edition.

Houston, K. (2013). *Shady Characters: The Secret Life of Punctuation, Symbols & Other Typographical Marks*. W.W. Norton & Company.

Sutcliffe, A. J. (1994). *The New York Public Library Writer's Guide to Style and Usage*. The New York Public Library and The Stonesong Press Inc.

◇ Kumaran Sathasivam
Writer and cosultant, Chennai, India
kumaran.sathasivam (at) gmail dot com

◇ S.K. Venkatesan
TNQ Books and Journals, Chennai, India
skvenkat (at) tnqsoftware dot co dot in

◇ Yakov Chandy
TNQ Books and Journals, Chennai, India
yakov (at) tnq dot co dot in