

These columns give a demonstration of PostScript Markup Hyphenation using the Tinydict. The left hand hyphenates by using the Cappella Archive Dividing Dictionary (4k) and the right hand the algorithms of F.M. Liang (73k) adopted by TeX. The PostScript hyphenation procedures were devised by Olavi Sakari and the Tinydict PostScript Typesetting Markup by David Byram-Wigfield.

See: <http://home.ricochet.com/osakari> for USA hyphenation

See: <http://www.cappella.demon.co.uk/tinyfiles/tinymenu.html/> for PostScript Markup

## DIGITAL WORD-DIVISION

Using PostScript Markup

### Using the Tiny Dividing Dictionary

Typeset pages are justified by spacing the words equally along the measure so that the last characters are in vertical alignment on the right. This is done by increasing the normal space unit of one-third of an 'M' slightly, or sometimes, less desirably, by reducing it. If the inter-word spaces remain unacceptably wide, the last word in the line may be split at a suitable point and the remainder carried over. A process known to some as word-division and by others called hyphenation.

The principles of word-division are very contentious and there are three subjective theories. The first divides words according to pronunciation; the second that the first half should give an expectation of what is to follow; whilst the third requires a pedantic separation according to etymology. So, you may find at-mosphere, atmos-phere, or atmo-sphere.

Traditional compositors have always disliked hyphens at the end of the first and last lines on the page; on the last line of a paragraph; and on more than two adjoining lines. The carrying-over of two letters was also frowned on, unless in narrow columns. These worthy intentions were often spoiled by the insertion of the wider quad space after every full point, producing 'holes' in an otherwise evenly spaced line of text.

Contemporary computer typesetting suffers from an equally unpleasant complaint whenever text is justified by 'tracking', whereby the distance between the individual characters of a word is also compressed or expanded. This produces a concertina fan-fold effect on adjacent lines and a restlessness on the page, in contrast to the calmer lines of fixed-width letterpress printing, where such inter-character kerning was only used to expand headlines and title-pages.

Automated word-division is incorporated into computer typesetting software and this can produce unexpected results, such as 'leg-end' for 'legend', and 'does-n't' is a commonplace typo in newspapers and magazines. One solution is to make the hyphenation procedures consult an exception list, before looking for the traditional break-points. The word 'Shake-speare' should be such an exception to prevent its appearance as 'Shakes-peare'.

When the justification process reaches the end of a line which would have unacceptably large spaces, it examines the carry-over word and looks for an existing hyphen to use as a break-point. If there isn't one, the exception list is searched for the word and, if not an exception, the procedure reads through a long list of hyphenation patterns to find a matching group of letters.

However, there are other complications. Punctuation marks and figures must be ignored; and all upper case characters temporarily converted to lower case before any exception and pattern search can begin. A 'threshold' value indicates the desirable, less desirable, and least desirable splitting places where hyphenation may occur. In printing parlance known as 'tight' or 'loose' setting.

The patterns are created by consulting authoritative word-division dictionaries such as the Oxford English Dictionary and Webster's. These hold in the region of 60,000 words and when grouped into similarities about ten thousand splitting patterns are found. Despite this large number, everyday text in normal prose only uses in the region of seven hundred combinations, with 'in-' and 'dis-' and '-ing' and '-ness' being obvious examples.

Odd numbers in the patterns select divisions in a threshold weighting order where 7 is weak and 1 is strong and these in turn are affected by the even numbers inserted to prevent any division. A narrow text column using threshold 1 would split ac-cord-ing-ly using the patterns (c1c) (6d3ing) (1ly.), but a wider book measure, using threshold 3, would divide as 'accord-ingly'. The even number 6 before 'ing' prevents words like 'mend-ing' or 'send-ing' being split incorrectly as 'men-ding' and 'sen-ding'.

Names of places and people are placed in the exceptions dictionary and splits are indicated by '9' to over-ride the lower threshold numbers 1, 3, 5, and 7 in the patterns. The exceptions (.re9al.) and (.re9ad.) will avoid a hyphen in those combinations of letters, which the prefix pattern (.re3) would otherwise create. Words like 'ghost' or 'through' inserted without 9s will never be split.

Most word-division dictionaries are jealously guarded by the commercial software designers, but the one freely available for computer typesetting was compiled by Frank M. Liang in 1983, whose PH.D thesis on the subject was published by Stanford University's Department of Computer Science. He developed an algorithm for most of the legitimate places to break words in roman languages.

These patterns were used by Professor Knuth for his TeX program for scientific typesetting and have been adapted by Olavi Sakari for PostScript Markup. Either the US or UK versions each consisting of some 8500 patterns may be inserted in the Tinydict. Do note that hyphenated text doubles interpreter time, so it is advisable to distil the file into the portable document format before printing.

The Cappella Archive Tinydivi dividing dictionary used in this column consists mainly of pre- and suffix combinations rather than many hundreds of short letter patterns. At 4k is it lives up to its name, and users may add their own patterns, which they are unable to do with the TeX dictionary.

### PostScript

PostScript was developed in 1982-5 by John Warnock and Chuck Geschke of Adobe Systems Inc. as a written description of the text and images to be printed on a page. The script provides instructions similar to those for finding buried treasure; a line is drawn from a starting point so many paces north, so many east, etc. The paces, measured at seventy-two printer's points to the inch, or minute fractions thereof are interpreted by a special chip placed in a laser printer.

PostScript rapidly became the universal page description language that it is today when desktop printing software, like PageMaker and Quark XPress, was developed to convert the images on the screen into scripted recipes. Despite their popularity for art-work and graphic design, editing a lengthy book over hundreds of pages on-screen is a daunting task.

There are other difficulties using commercial typesetting programs. The generated PostScript scripts are impossible to edit without a return to the original software version, which makes archiving unreliable. Even worse, adjoining words and characters in the text are constantly moved closer or further apart in an attempt to 'improve' the appearance of the text on the page.

This typographical restlessness involves a software-generated combination of 'tracking' and 'kerning' with the movement coordinates often expressed to six decimal places (i.e. one millionth of an inch).

The complex digital barrier between the computer screen and the printer, makes most users unaware of the elegance, accuracy and efficiency of PostScript as a scripted printing language; requiring as it does only the simplest of text editors to send typesetting instructions directly to the printer interpreter.

See the E-book Hyphenation from The Hyphenologist:  
<http://www.hyphenologist.co.uk>

### TeX hyphenation patterns

Typeset pages are justified by spacing the words equally along the measure so that the last characters are in vertical alignment on the right. This is done by increasing the normal space unit of one-third of an 'M' slightly, or sometimes, less desirably, by reducing it. If the inter-word spaces remain unacceptably wide, the last word in the line may be split at a suitable point and the remainder carried over. A process known to some as word-division and by others called hyphenation.

The principles of word-division are very contentious and there are three subjective theories. The first divides words according to pronunciation; the second that the first half should give an expectation of what is to follow; whilst the third requires a pedantic separation according to etymology. So, you may find at-mosphere, atmos-phere, or atmo-sphere.

Traditional compositors have always disliked hyphens at the end of the first and last lines on the page; on the last line of a paragraph; and on more than two adjoining lines. The carrying-over of two letters was also frowned on, unless in narrow columns. These worthy intentions were often spoiled by the insertion of the wider quad space after every full point, producing 'holes' in an otherwise evenly spaced line of text.

Contemporary computer typesetting suffers from an equally unpleasant complaint whenever text is justified by 'tracking', whereby the distance between the individual characters of a word is also compressed or expanded. This produces a concertina fan-fold effect on adjacent lines and a restlessness on the page, in contrast to the calmer lines of fixed-width letterpress printing, where such inter-character kerning was only used to expand headlines and title-pages.

Automated word-division is incorporated into computer typesetting software and this can produce unexpected results, such as 'leg-end' for 'legend', and 'does-n't' is a commonplace typo in newspapers and magazines. One solution is to make the hyphenation procedures consult an exception list, before looking for the traditional break-points. The word 'Shake-speare' should be such an exception to prevent its appearance as 'Shakes-peare'.

When the justification process reaches the end of a line which would have unacceptably large spaces, it examines the carry-over word and looks for an existing hyphen to use as a break-point. If there isn't one, the exception list is searched for the word and, if not an exception, the procedure reads through a long list of hyphenation patterns to find a matching group of letters.

However, there are other complications. Punctuation marks and figures must be ignored; and all upper case characters temporarily converted to lower case before any exception and pattern search can begin. A 'threshold' value indicates the desirable, less desirable, and least desirable splitting places where hyphenation may occur. In printing parlance known as 'tight' or 'loose' setting.

The patterns are created by consulting authoritative word-division dictionaries such as the Oxford English Dictionary and Webster's. These hold in the region of 60,000 words and when grouped into similarities about ten thousand splitting patterns are found. Despite this large number, everyday text in normal prose only uses in the region of seven hundred combinations, with 'in-' and 'dis-' and '-ing' and '-ness' being obvious examples.

Odd numbers in the patterns select divisions in a threshold weighting order where 7 is weak and 1 is strong and these in turn are affected by the even numbers inserted to prevent any division. A narrow text column using threshold 1 would split ac-cord-ing-ly using the patterns (c1c) (6d3ing) (1ly.), but a wider book measure, using threshold 3, would divide as 'accord-ingly'. The even number 6 before 'ing' prevents words like 'mend-ing' or 'send-ing' being split incorrectly as 'men-ding' and 'sen-ding'.

Names of places and people are placed in the exceptions dictionary and splits are indicated by '9' to over-ride the lower threshold numbers 1, 3, 5, and 7 in the patterns. The exceptions (.re9al.) and (.re9ad.) will avoid a hyphen in those combinations of letters, which the prefix pattern (.re3) would otherwise create. Words like 'ghost' or 'through' inserted without 9s will never be split.

Most word-division dictionaries are jealously guarded by the commercial software designers, but the one freely available for computer typesetting was compiled by Frank M. Liang in 1983, whose PH.D thesis on the subject was published by Stanford University's Department of Computer Science. He developed an algorithm for most of the legitimate places to break words in roman languages.

These patterns were used by Professor Knuth for his TeX program for scientific typesetting and have been adapted by Olavi Sakari for PostScript Markup. Either the US or UK versions each consisting of some 8500 patterns may be inserted in the Tinydict. Do note that hyphenated text doubles interpreter time, so it is advisable to distil the file into the portable document format before printing.

The Cappella Archive Tinydivi dividing dictionary used on the left consists mainly of pre- and suffix combinations rather than many hundreds of short letter patterns. At 4k is it lives up to its name, and users may add their own patterns, which they are unable to do with the TeX dictionary.

### PostScript

PostScript was developed in 1982-5 by John Warnock and Chuck Geschke of Adobe Systems Inc. as a written description of the text and images to be printed on a page. The script provides instructions similar to those for finding buried treasure; a line is drawn from a starting point so many paces north, so many east, etc. The paces, measured at seventy-two printer's points to the inch, or minute fractions thereof are interpreted by a special chip placed in a laser printer.

PostScript rapidly became the universal page description language that it is today when desktop printing software, like PageMaker and Quark XPress, was developed to convert the images on the screen into scripted recipes. Despite their popularity for art-work and graphic design, editing a lengthy book over hundreds of pages on-screen is a daunting task.

There are other difficulties using commercial typesetting programs. The generated PostScript scripts are impossible to edit without a return to the original software version, which makes archiving unreliable. Even worse, adjoining words and characters in the text are constantly moved closer or further apart in an attempt to 'improve' the appearance of the text on the page.

This typographical restlessness involves a software-generated combination of 'tracking' and 'kerning' with the movement coordinates often expressed to six decimal places (i.e. one millionth of an inch).

The complex digital barrier between the computer screen and the printer, makes most users unaware of the elegance, accuracy and efficiency of PostScript as a scripted printing language; requiring as it does only the simplest of text editors to send typesetting instructions directly to the printer interpreter.