**WWN Development Document**

Word HTML Bugs :

bugs, bug-fixes, and problems

Word’s Navigation pane shows the table-of-contents (View : Show : Navigation pane).

* Contents:
  + For Word HTML bugs
    - Bugs, bug-fixes, and HTML editing
  + Shortcomings in Word’s HTML, due to document formatting and layout

This document was created by the WWN author for his own use in developing WWN. It is included in the WWN repo, as other developers may find it useful.

# Fixing and editing Word HTML

## Word TOC HTML

* TOC info is also in: word-html--table-of-contents.docx

### Summary

* Summary
  + The objective: change link formatting (underline, color change), extract html to put in a TOC div
  + The changes required are to add CSS class specifications... need to describe this
  + The general format of the TOC-entry paragraphs is shown below.
* General format of TOC-entry paragraphs with hyperlinks
  + The paragraph has one child tag, and that child tag can have children

<p class=MsoToc1>[nested span opening-tags]<a href="#\_Toc68878247">[span lang= opening-tag][text][span lang= closing-tag][page-number span tags, with display:none;]</a>[nested span closing-tags]

<p class=MsoToc1>[anchor tags with no text]<a href="#\_Toc68878247">[text]</a>[nested span closing-tags]

* + All TOC-entry paragraphs start with <p class=MsoToc[1-9]>
  + Not shown: there can potentially be whitespace after opening-tags and/or before closing-tags, e.g., due to newlines.
  + [nested span opening-tags]
    - Can be none, one or two of these span tags
    - Two types of span opening-tags:
      * <span lang=DE>
      * <span class=MsoHyperlink>
    - If one, I've just seen <span class=MsoHyperlink>
    - If both, they are nested and the types can be in any order

<span lang=DE><span class=MsoHyperlink>...</span></span>

<span class=MsoHyperlink><span lang=DE>...</span></span>

* + - If one or both, they contain the <a>...</a> tag
      * The span closing-tags go after the </a>: [nested span closing-tags]\*
* General format of TOC-entry paragraphs without hyperlinks:
  + See below

### Survey of the types of TOC entries in the Word HTML:

#### HTML TOC entries, based on the possible TOC options

* HTML TOC entries, based on the possible TOC options
  + For a particular TOC, all TOC entries are of the same type (same HTML elements)
* **TOC options: hyperlinks, TOC includes no page numbers:**

<p class=MsoToc1><span class=MsoHyperlink><a href="#\_Toc68878247">1 L1</a></span></p>

* **TOC options: unknown**
  + The TOC entry has an anchor element (<a>), but **NOT** "<span class=MsoHyperlink>"
  + This TOC entry has a hyperlink, but I can't figure-out what creates this type of TOC entry
    - It may be a combination of: the headings used, other document content, and TOC options
    - The TOC options appear to be: hyperlinks, page-numbers right-aligned
  + Example: (headings-and-bullets.docx, headings-and-bullets.htm), (sys-admin.docx, sys-admin.htm)

<p class=MsoToc1><a href="#\_Toc67590875">1 Level 1<span style='color:windowtext;

display:none;text-decoration:none'>. </span><span

style='color:windowtext;display:none;text-decoration:none'>1</span></a></p>

* TOC options: hyperlinks, TOC includes page numbers
  + Note:
    - When page-numbers are included, there are two <span> elements after the TOC-entry's text.
    - Those <span> elements' text does not get displayed, and can be ignored.
      * Due to style='...display:none;...'
    - These <span> elements may be created by MS Word for its own use in opening the \*.htm file, and displaying the TOC.
  + TOC options: hyperlinks, TOC includes page numbers, right-align page numbers

<p class=MsoToc1><span class=MsoHyperlink><a href="#\_Toc68878729">1 L1<span

style='color:windowtext;display:none;text-decoration:none'>. </span><span

style='color:windowtext;display:none;text-decoration:none'>1</span></a></span></p>

* + TOC options: hyperlinks, TOC includes page numbers, page numbers not right-aligned

<p class=MsoToc1><span class=MsoHyperlink><a href="#\_Toc68879103">1 L1<span

style='color:windowtext;display:none;text-decoration:none'> </span><span

style='color:windowtext;display:none;text-decoration:none'>1</span></a></span></p>

* TOC options: no hyperlinks,
  + TOC options: no hyperlinks, no page numbers

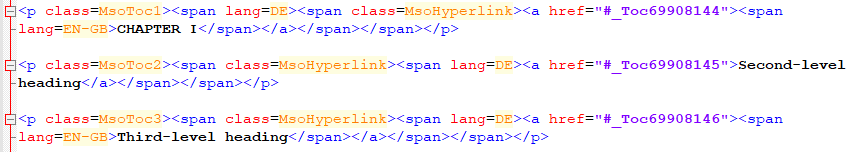
<p class=MsoToc1>1 L1</p>

* + TOC options: no hyperlinks, page numbers

<p class=MsoToc1>1 L1................................................................................................................................................. 1</p>

#### TOC paragraph with one or more tags: "<span lang= ..."

* **TOC paragraph with one or more tags: "<span lang= ..."**
  + "<span lang= ..." in 4 places below



#### TOC paragraph with multiple anchor tags: "<a..."

<p class="MsoToc1"><a class="tocAnchor" name="\_Toc74303531"></a><a name="\_Toc74303626"></a><a name="\_Toc74304313"></a><a name="\_Toc74304544"></a><a href="#\_Toc74313249">1 L1-1</a></p>

### Experiments

#### Overview

* This section records experiments I did to determine what HTML Word creates for TOC entries.
* Scope of experiments
  + Normal.dot files used:
    - Used my file, with my Word customizations
    - Used MS Word's default file
  + TOC options used
    - Options that varied across experiments
      * With and without: page numbers, hyperlinks
      * Print and web layouts
    - Options used in all experiments
      * Format: from template
      * Options: used default
      * Modify: used default
      * Save as: Web, filtered
  + Experiment files are in my directory: Word-TOC-entry-HTML-types

#### Testing the different TOC-Option combinations

* Word new
  + Show page numbers: off
    - Use hyperlinks: off
      * Web layout (v1)

<p class=MsoToc1>1 L1</p>

* + - * Print layout (v2)

<p class=MsoToc1>1 L1</p>

* + - Use hyperlinks: on
      * Web layout (v3)

<p class=MsoToc1><span class=MsoHyperlink><a href="#\_Toc68878247">1 L1</a></span></p>

* + - * Print layout (v7)

<p class=MsoToc1><span class=MsoHyperlink><a href="#\_Toc68879418">1 L1</a></span></p>

* + Show page numbers: on
    - Use hyperlinks: off
      * Web layout (v4)

<p class=MsoToc1>1 L1................................................................................................................................................. 1</p>

* + - Use hyperlinks: on
      * Web layout
        + Right align page numbers: on (v5)

<p class=MsoToc1><span class=MsoHyperlink><a href="#\_Toc68878729">1 L1<span

style='color:windowtext;display:none;text-decoration:none'>. </span><span

style='color:windowtext;display:none;text-decoration:none'>1</span></a></span></p>

* + - * + Right align page numbers: off (v6)

<p class=MsoToc1><span class=MsoHyperlink><a href="#\_Toc68879103">1 L1<span

style='color:windowtext;display:none;text-decoration:none'> </span><span

style='color:windowtext;display:none;text-decoration:none'>1</span></a></span></p>

* XYPlorer new
  + Show page numbers: off
    - Use hyperlinks: off
      * Web layout (v)
      * Print layout (v)
    - Use hyperlinks: on
      * Web layout (v)

<p class=MsoToc1><span class=MsoHyperlink><a href="#\_Toc68899583">1 L1</a></span></p>

* + - * Print layout (v)
  + Show page numbers: on
    - Use hyperlinks: off
      * Web layout (v)
    - Use hyperlinks: on
      * Web layout
        + Right align page numbers: on (v)

<p class=MsoToc1><span class=MsoHyperlink><a href="#\_Toc68899914">1 L1<span

Print layout (v10)

<p class=MsoToc1><span class=MsoHyperlink><a href="#\_Toc68901825">1 L1<span

* + - * + Right align page numbers: off (v)

#### TOC-experiment-01.docx

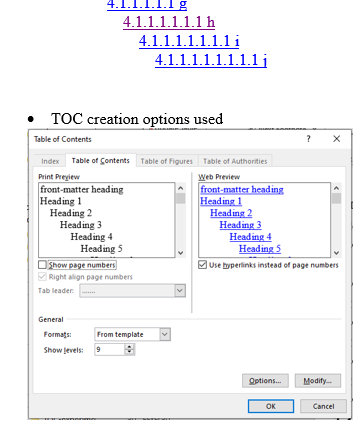
* Summary
  + TOC entries have links
  + HTML TOC entries

<p class=MsoToc2><span class=MsoHyperlink><a href="#\_Toc68857208">1.2 Level 1.2</a></span></p>

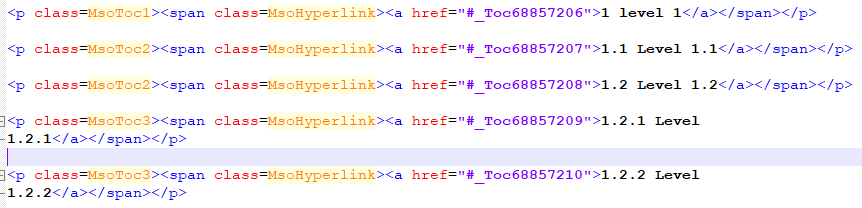
<p class=MsoToc3><span class=MsoHyperlink><a href="#\_Toc68857209">1.2.1 Level

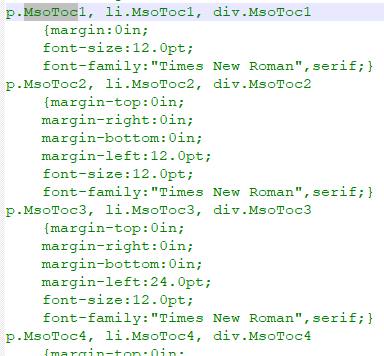
1.2.1</a></span></p>

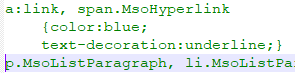
* Config



* HTML





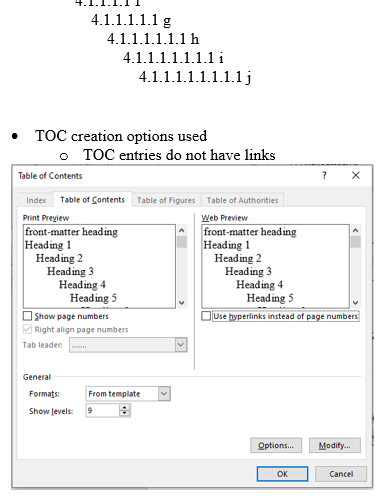


#### TOC-experiment-02.docx

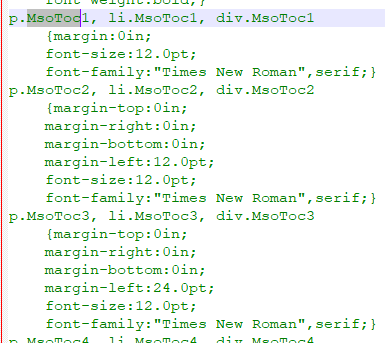
* Summary
  + TOC entries do not have links.
  + HTML TOC entries

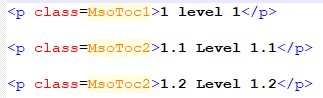
<p class=MsoToc1>1 level 1</p>

* Config

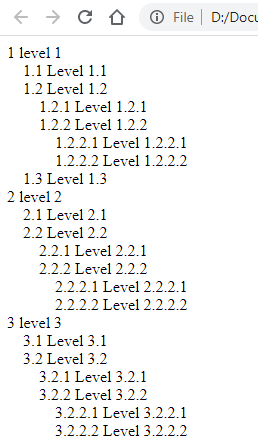


* HTML



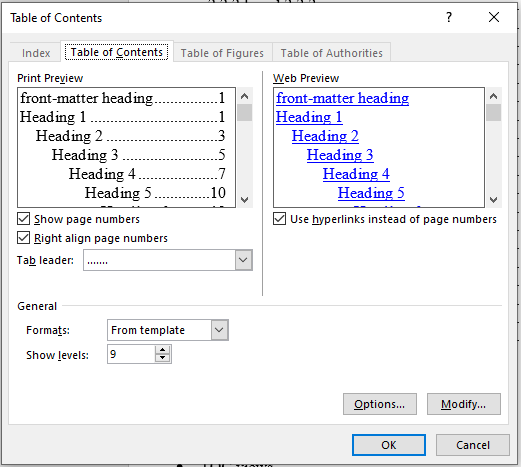


* Chrome
  + No hyperlinks

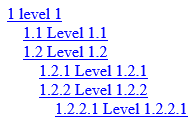


#### TOC-experiment-03.docx

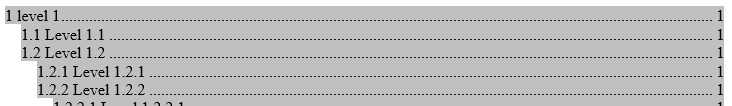
* Summar
  + "Show page numbers" results in extra HTML, which is not displayed
  + The Web-Layout and Print-Layout views differ
    - Note this in user manual
* TOC creation options used



* TOC views
  + Web Layout



* + Print Layout

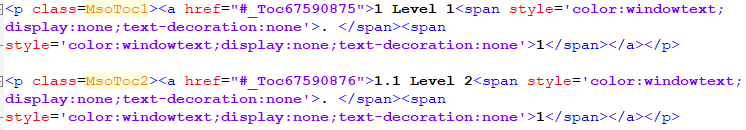


* HTML



#### Word doc created via File : New

* Other tests may have created the Word doc via "File:New" or the XYPlorer new-item Word doc
* Test 1



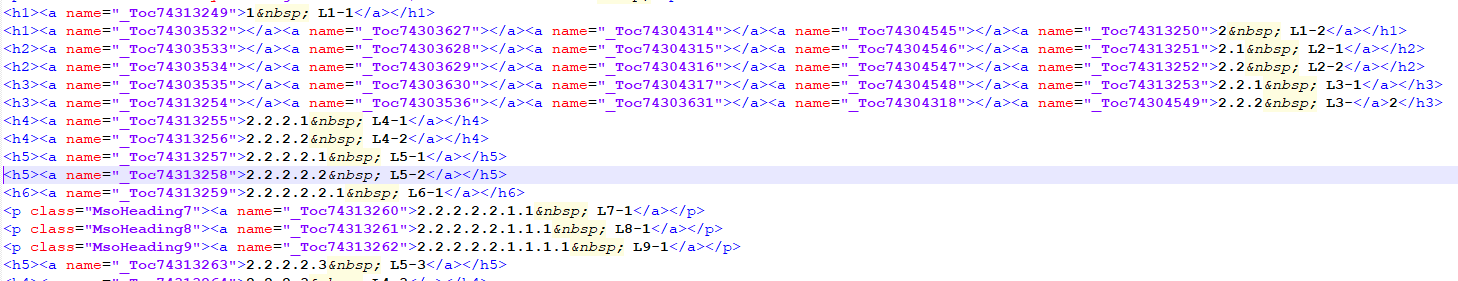
* Test 2



* Test 3
  + Notes
    - Multiple names created for a header, and TOC has multiple href's to multiple names
    - Multiple names probably due to repeatedly creating and deleting a TOC in the doc
    - toc-test--.docx
  + TOC entry

<p class="MsoToc1"><a class="tocAnchor" name="\_Toc74303531"></a><a name="\_Toc74303626"></a><a name="\_Toc74304313"></a><a name="\_Toc74304544"></a><a href="#\_Toc74313249">1 L1-1</a></p>

* + Heading entries

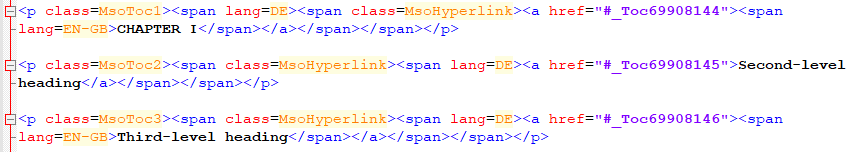


#### Word's default Normal.dot file



#### example-book-formatting.doc

* The file is .doc, from the Internet
* "<span lang= ..." in 4 places here



## Word unordered lists

### Summary

* <p class=[class-name] ...>, possible class names:

MsoListParagraphCxSpFirst

MsoListParagraphCxSpMiddle

MsoListParagraphCxSpLast

MsoListParagraph

MsoNormal

--Found later during testing of downloaded Word docs

MsoBodyText

ScrollListBullet (not supported in my code, needs more investigation)

O-BodyText (not supported in my code, needs more investigation)

<p class=ScrollListBullet><span style='font-family:Symbol'>·<span

style='font:7.0pt "Times New Roman"'>&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;

</span></span>Step 3 allows you to choose the installation folder.  The default

is <i>C:\Program Files\NYSPRL</i>.</p>

* <p... style=[style\_spec]>

<p... style='margin-left:[.\*]in;text-indent:[.\*]in'>

* + [style\_spec] usually has "margin-left:[.\*]in;", but not always
    - If omitted, it's defined in the Word <style> section, for the paragraph class
  + Indentation seems ok and consistent
  + <p class=ScrollListBullet>
    - p tag does not have other attributes, e.g., no style=. It's the only bullet like that
* <p ...><span style='font-family:[font\_name]'>[bullet\_symbol]<span style=>[&nbsp;]</span></span>
  + outter span can be preceded by, and enclosed by: <span class=MsoHyperlink>
  + outter span can be preceded by, and enclosed by: <font size=2 face=Symbol>
  + outer span:
    - style can also have "font-size:10.0pt;", "color:windowtext;", "text-decoration:none"
    - can have: lang=EN-GB
  + 3 font symbols
    - <span style='font-family:Wingdings'>§
    - <span style='font-family:Symbol'>·
    - <span style='font-family:"Courier New"'>o
  + inner span:
    - "&nbsp;"'s can have a space after them, before </span>
* Algorithm for list-item detection
  + paragraph has expected class-name, and style with "text-indent:"
  + Get strings for paragraph
  + First string:
    - a bullet-symbol
    - enclosed in a span, with sytle and expected font-family
  + Second string
    - "&nbsp;"'s and possibly whitespace
    - enclosed in a span with style
    - the span's parent is in the bullet-symbol's parent

### Survey of the types of unordered lists

#### From my docs

* Symbol types found

<p class=MsoListParagraphCxSpMiddle style='margin-left:1.25in;text-indent:-.25in'>

<span style='font-family:Wingdings'>§<span style='font:7.0pt "Times New Roman"'>&nbsp;

</span></span>

[text]</p>

<p class=MsoListParagraphCxSpFirst style='margin-left:.25in;text-indent:-.25in'>

<span style='font-family:Symbol'>·<span style='font:7.0pt "Times New Roman"'>&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;

</span></span>

[text]</p>

<p class=MsoListParagraphCxSpMiddle style='margin-left:.75in;text-indent:-.25in'>

<span style='font-family:"Courier New"'>o<span style='font:7.0pt "Times New Roman"'>&nbsp;&nbsp;

</span></span>

[text]</p>

#### Downloaded doc

<p class=MsoNormal style='margin-left:.5in;text-indent:-.25in'>

<span lang=EN-GB style='font-family:Symbol'>·<span style='font:7.0pt "Times New Roman"'>&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;

</span></span>

<span lang=EN-GB>

[text]</p>

### Unordered lists: my Word app

#### Word HTML bullet symbols

##### Wingdings'>§<

<p class=MsoListParagraphCxSpMiddle style='margin-left:1.25in;text-indent:-.25in'><span

style='font-family:Wingdings'>§<span style='font:7.0pt "Times New Roman"'>&nbsp;

</span></span>level 3</p>

<p class=MsoListParagraphCxSpMiddle style='margin-left:1.25in;text-indent:-.25in'><span

style='font-size:10.0pt;font-family:Wingdings'>§<span style='font:7.0pt "Times New Roman"'>&nbsp;

</span></span>  <span style='font-size:10.0pt;font-family:"Courier New"'>pip

install pprintpp </span></p>

>>> Space after &nbsp;

>>> Inside <span class=MsoHyperlink>

>>> "Wingdings;" followed by: color:windowtext;text-decoration:none

<p class=MsoListParagraphCxSpMiddle style='margin-left:1.25in;text-indent:-.25in'><span

class=MsoHyperlink><span style='font-family:Wingdings;color:windowtext;

text-decoration:none'>§<span style='font:7.0pt "Times New Roman"'>&nbsp; </span></span></span><a

href="https://pymotw.com/2/ConfigParser/">https://pymotw.com/2/ConfigParser/</a></p>

<p class=MsoListParagraphCxSpMiddle style='margin-left:1.25in;text-indent:-.25in'><span

class=MsoHyperlink><span style='font-family:Wingdings;text-decoration:none'>§<span

style='font:7.0pt "Times New Roman"'>&nbsp; </span></span></span><a

href="https://docs.microsoft.com/en-us/windows-server/administration/windows-commands/chkdsk">https://docs.microsoft.com/en-us/windows-server/administration/windows-commands/chkdsk</a></p>

>>> full html: D:\Documents\Professional-projects\My-web-site-development\Word-to-HTML\Word-to-HTML-experiments\headings-and-bullets--full-html.htm

<p class=MsoListParagraphCxSpMiddle style='margin-left:1.25in;mso-add-space:

auto;text-indent:-.25in;mso-list:l4 level3 lfo41'><![if !supportLists]><span

style='font-family:Wingdings;mso-fareast-font-family:Wingdings;mso-bidi-font-family:

Wingdings'><span style='mso-list:Ignore'>§<span style='font:7.0pt "Times New Roman"'>&nbsp;

</span></span></span><![endif]>level 3</p>

###### Found in testing, had to fix my code to handle this

<p class=MsoListParagraphCxSpMiddle style='margin-left:1.25in;text-indent:-.25in'><span

class=MsoHyperlink><span style='font-family:Wingdings;text-decoration:none'>§<span

style='font:7.0pt "Times New Roman"'>&nbsp; </span></span></span><a

href="https://docs.microsoft.com/en-us/windows-server/administration/windows-commands/chkdsk">https://docs.microsoft.com/en-us/windows-server/administration/windows-commands/chkdsk</a></p>

>>> "class=MsoListParagraph"

<p class=MsoListParagraph style='margin-left:1.25in;text-indent:-.25in'><span

style='font-family:Wingdings'>§<span style='font:7.0pt "Times New Roman"'>&nbsp;

</span></span>Click &quot;Expand All&quot; type of link to see all downloads</p>

<p class=MsoListParagraphCxSpMiddle style='margin-left:1.25in;text-indent:-.25in'><span

style='font-size:10.0pt;font-family:Wingdings'>§<span style='font:7.0pt "Times New Roman"'>&nbsp;

</span></span> <span style='font-size:10.0pt;font-family:"Courier New"'>127.0.0.1

1www.jimyuill.org</span></p>

##### >·<

<p class=MsoListParagraphCxSpFirst style='margin-left:.25in;text-indent:-.25in'><span

style='font-family:Symbol'>·<span style='font:7.0pt "Times New Roman"'>&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;

</span></span>YouTube video shows how to install new ssd and hdd</p>

* Note: p style doesn't have "margin-left:.25in;"

<p class=MsoListParagraphCxSpFirst style='text-indent:-.25in'><span

style='font-family:Symbol'>·<span style='font:7.0pt "Times New Roman"'>&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;

</span></span>level 1</p>

##### Courier New"'>o<

<p class=MsoListParagraphCxSpMiddle style='margin-left:.75in;text-indent:-.25in'><span

style='font-family:"Courier New"'>o<span style='font:7.0pt "Times New Roman"'>&nbsp;&nbsp;

</span></span>level 2</p>

<p class=MsoListParagraphCxSpMiddle style='margin-left:.75in;text-indent:-.25in'><span

style='font-family:"Courier New"'>o<span style='font:7.0pt "Times New Roman"'>&nbsp;&nbsp;

</span></span><a href="https://www.youtube.com/watch?v=uEIlmHyJ8V4">https://www.youtube.com/watch?v=uEIlmHyJ8V4</a>

</p>

<p class=MsoListParagraphCxSpMiddle style='margin-left:.75in;text-indent:-.25in'><span

style='font-size:10.0pt;font-family:"Courier New"'>o<span style='font:7.0pt "Times New Roman"'>&nbsp;&nbsp;&nbsp;

</span></span><span style='font-size:10.0pt;font-family:"Courier New"'>pip

install yamllint</span></p>

#### Indentation

* Summary
  + Indentation seems ok and consistent
* From: headings-and-bullets-01.htm
* List starts not indented, it's created by my VBA macro F11

L1: 'margin-left:.25in;text-indent:-.25in'

L2: 'margin-left:.75in;text-indent:-.25in'

L3: 'margin-left:1.25in;text-indent:-.25in'

L4: 'margin-left:1.75in;text-indent:-.25in'

L5: 'margin-left:2.25in;text-indent:-.25in'

L6: 'margin-left:2.75in;text-indent:-.25in'

L7: 'margin-left:3.25in;text-indent:-.25in'

L8: 'margin-left:3.75in;text-indent:-.25in'

L9: 'margin-left:4.25in;text-indent:-.25in'

* This list starts indented, it's created by default

L1: 'text-indent:-.25in'

L2: 'margin-left:1.0in;text-indent:-.25in'

L3: 'margin-left:1.5in;text-indent:-.25in'

L4: 'margin-left:2.0in;text-indent:-.25in'

L5: 'margin-left:2.5in;text-indent:-.25in'

L6: 'margin-left:3.0in;text-indent:-.25in'

L7: 'margin-left:3.5in;text-indent:-.25in'

L8: 'margin-left:4.0in;text-indent:-.25in'

L9: 'margin-left:4.5in;text-indent:-.25in'

### Unordered lists: downloaded .doc file

* example-book-formatting.doc
  + Deeper levels are not possible (not supported by older Word?)
  + level-1:

<p class=MsoNormal style='margin-left:.5in;text-indent:-.25in'><span

lang=EN-GB style='font-family:Symbol'>·<span style='font:7.0pt "Times New Roman"'>&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;

</span></span><span lang=EN-GB>Id facilis reformidans eum</span></p>

### Other unordered lists

* MS-tutorial--Deep Learning for Signal and Information Processing.docx

<p class=MsoBodyText style='margin-left:26.25pt;text-indent:-26.25pt'><span

style='font-family:Symbol'>·<span style='font:7.0pt "Times New Roman"'>&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;

</span></span><b><i>Definition 1</i></b><i>:</i> A class of machine learning ...

</p>

* pesticide-report.htm
  + The ScrollListBullet's may be spurious; apparently not useful text in them?
  + ScrollListBullet3 does not have a bullet symbol

<p class=ScrollListBullet3 style='margin-left:.75in;text-indent:0in'>&nbsp;</p>

<p class=ScrollListBullet><span style='font-family:Symbol'>·<span

style='font:7.0pt "Times New Roman"'>&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;

</span></span>&nbsp;</p>

* software-install-guide.html
  + ScrollListBullet
    - <p...> has no text-indent
  + ScrollListBullet3
    - seems to be spurious and is just used to hold a table

<p class=ScrollListBullet3 style='margin-left:.75in;text-indent:0in'>

<p class=ScrollListBullet><span style='font-family:Symbol'>·<span

style='font:7.0pt "Times New Roman"'>&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;

</span></span>Step 3 allows you to choose the installation folder.  The default

is <i>C:\Program Files\NYSPRL</i>.</p>

* media-use-terms.html

<p class=O-BodyText style='margin-left:.75in;text-align:justify;text-indent:

-.25in;line-height:normal'><font size=2 face=Symbol><span lang=EN

style='font-size:11.0pt;font-family:Symbol'>·<font size=1 face="Times New Roman"><span

style='font:7.0pt "Times New Roman"'>&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp; </span></font></span></font><font

size=2 face=Calibri><span lang=EN style='font-size:11.0pt;font-family:"Calibri",sans-serif'>in

logos, trademarks, services marks or any other branding or identifiers.</span></font></p>

## Word ordered-lists

* An example of a typical list-item paragraph:
  + The line labels (e.g., Line A:) are not in the HTML

Line A: <p class=MsoListParagraphCxSpMiddle style='margin-left:1.5in;text-indent:-1.5in'>

Line B: <span style='font:7.0pt "Times New Roman"'>&nbsp;&nbsp;&nbsp;&nbsp;</span>i.<span

Line C: style='font:7.0pt "Times New Roman"'>&nbsp;&nbsp;</span>List-item text</p>

* There are two commonly-found bugs in these list-items:
  + The list-item symbol is often not properly indented
  + The text after the list-item symbol is often not properly indented
    - (it does not have the proper number of spaces between the symbol and the start of the text)

### Summary of ordered-list list-items’ HTML

* <p class=[class-name] ...>
  + Possible class names:

MsoListParagraphCxSpFirst

MsoListParagraphCxSpMiddle

MsoListParagraphCxSpLast

MsoNormal

* + class-name NOT observed, but used with unordered lists

MsoListParagraph

---- Found in later testing

MsoBodyText

* <p... style=...> starts with:
  + always has style with "text-indent:"
    - The 'text-indent: ..." tag is inconsistent and often incorrect
    - Using 'text-indent:-.25in' for all levels fixes the problem
  + style usually has "margin-left:", but not always

<p class=MsoListParagraphCxSpFirst style='text-indent:-.25in'>

<p class=MsoListParagraphCxSpLast style='margin-left:4.0in;text-indent:-.25in'>

<p class=MsoListParagraphCxSpMiddle style='margin-left:1.0in;text-indent:-.25in'>

<p class=MsoListParagraphCxSpLast style='margin-left:4.0in;text-indent:-.25in'>

<p class=MsoNormal style='margin-left:.5in;text-indent:-.25in'>

* Summary: headings-bullets-default-norma-dot.htm

<p...>[item-number"."]<span style=>["&nbsp;]</span>[text]

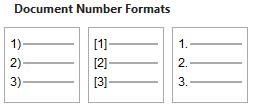
<p...><span style=>["&nbsp;]</span>[item-number"."]<span style=>["&nbsp;]</span>[text]

* Summary: multi-level lists, from downloaded .doc file

<p...><span lang= >[item\_number"."]<span style=>["&nbsp;"s]</span></span>

<p...><span lang= ><span style=>[&nbsp;]</span>[item\_number"."]<span style=>[&nbsp;]</span></span>

* Item-number formats
  + From Word ribbon:
    - "item-number" can be followed by: ), ], .
    - "item-number" can be preceded by: [



* MS-tutorial--Deep Learning for Signal and Information Processing.docx
  + <p...>[item-number")"]<span

<p class=MsoBodyText style='margin-left:.5in;text-indent:-.25in'>1)<span

* headings-and-bullets.docx

<p class="MsoListParagraphCxSpFirst" style="text-indent:-.25in">1)<span style='font:7.0pt "Times New Roman"'>&nbsp;&nbsp;&nbsp;&nbsp;&nbsp; </span>A</p>

<p class="MsoListParagraphCxSpFirst" style="text-indent:-.25in">1.<span style='font:7.0pt "Times New Roman"'>&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;</span>One</p>

<p class="MsoListParagraphCxSpFirst" style="text-indent:-.25in">[1]<span style='font:7.0pt "Times New Roman"'>&nbsp;&nbsp; </span>First</p>

### Survey

#### Ordered-lists: my Word app

##### Indentation problems

* Summary
  + The 'text-indent: ..." tag is inconsistent and often incorrect
  + Using 'text-indent:-.25in' for all levels fixes the problem
* From: headings-and-bullets-01.htm

l-1: 'text-indent:-.25in'

l-2: 'margin-left:1.0in;text-indent:-.25in'

l-3: 'margin-left:1.5in;text-indent:-1.5in'

l-4: 'margin-left:2.0in;text-indent:-.25in'

l-5: 'margin-left:2.5in;text-indent:-.25in'

l-6: 'margin-left:3.0in;text-indent:-3.0in'

l-7: 'margin-left:3.5in;text-indent:-.25in'

l-8: 'margin-left:4.0in;text-indent:-.25in'

l-9: 'margin-left:4.5in;text-indent:-4.5in'

#### headings-bullets-default-norma-dot.htm

* Summary: headings-bullets-default-norma-dot.htm

<p...>[item-number"."]<span style=>["&nbsp;]</span>[text]

<p...><span style=>["&nbsp;]</span>[item-number"."]<span style=>["&nbsp;]</span>[text]

<p class=MsoListParagraphCxSpFirst style='text-indent:-.25in'>1.<span

style='font:7.0pt "Times New Roman"'>&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp; </span>L-1</p>

<p class=MsoListParagraphCxSpMiddle style='margin-left:1.0in;text-indent:-.25in'>a.<span

style='font:7.0pt "Times New Roman"'>&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp; </span>L-2</p>

<p class=MsoListParagraphCxSpMiddle style='margin-left:1.5in;text-indent:-1.5in'><span

style='font:7.0pt "Times New Roman"'>&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;

</span>i.<span style='font:7.0pt "Times New Roman"'>&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;

</span>L-3</p>

<p class=MsoListParagraphCxSpLast style='margin-left:4.0in;text-indent:-.25in'>b.<span

style='font:7.0pt "Times New Roman"'>&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp; </span>L-8</p>

#### Ordered-lists: downloaded .doc file

* Summary:

<p...><span lang= >[item-number"."]<span style=>["&nbsp;]</span></span>

* example-book-formatting.doc
  + Deeper levels are not possible (not supported by older Word?)
  + level-1:

<p class=MsoNormal style='margin-left:.5in;text-indent:-.25in'><span

lang=EN-GB>1.<span style='font:7.0pt "Times New Roman"'>&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;

</span></span><span lang=EN-GB>Id facilis reformidans eum</span></p>

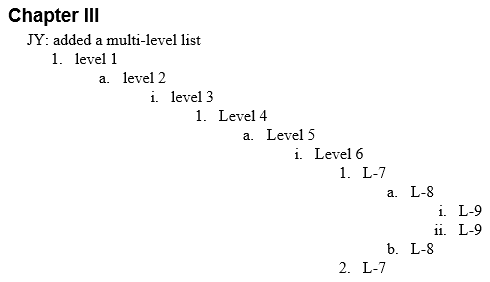
#### Word multi-level lists

##### downloaded .doc file

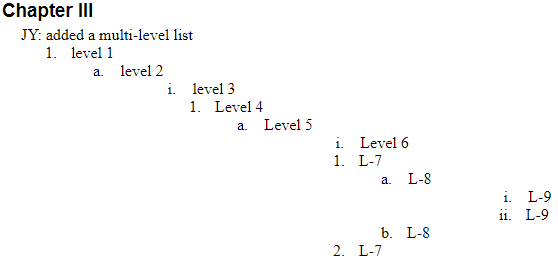
* Summary
  + **Formatting problems**, worse in lower levels
  + Bullets' HTML start with:

<p class=MsoNormal style='margin-left:50.2pt;text-indent:-.25in'>

* + Experiments to fix it, not successful
    - Didn't change bullet HTML starting element <p... :
      * Saving as .docx
      * Copying original contents to a new .docx
* example-book-formatting--multi-level-list-added.doc
  + I added multi-level list to: example-book-formatting.doc
* example-book-formatting--multi-level-list-added--v2.docx
  + I created this from .doc versioin
* Doc:



* WHWN HTML



###### Generalized HTML

* Summary
  + [item\_number"."] and <span> are siblings

<p...><span lang= >[item\_number"."]<span style=>["&nbsp;"s]</span></span>

<p...><span lang= ><span style=>[&nbsp;]</span>[item\_number"."]<span style=>[&nbsp;]</span></span>

<p... ><span lang=EN-GB>1.<span style='font:7.0pt "Times New Roman"'>&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;</span></span>

<p... ><span lang=EN-GB><span style='font:7.0pt "Times New Roman"'>&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;

</span>i.<span style='font:7.0pt "Times New Roman"'>&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;

</span></span>

###### Examples

* Summary: from downloaded .doc file

<p...><span lang= >[item\_number"."]<span style=>["&nbsp;]</span></span>

<p...><span lang= ><span style=... >["&nbsp;]</span>[item\_number"."]<span>[&nbsp;]</span></span>

<p class=MsoNormal style='margin-left:50.2pt;text-indent:-.25in'><span

lang=EN-GB>1.<span style='font:7.0pt "Times New Roman"'>&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;

</span></span><span lang=EN-GB>level 1</span></p>

<p class=MsoNormal style='margin-left:86.2pt;text-indent:-.25in'><span

lang=EN-GB>a.<span style='font:7.0pt "Times New Roman"'>&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;

</span></span><span lang=EN-GB>level 2</span></p>

<p class=MsoNormal style='margin-left:122.2pt;text-indent:-122.2pt'><span

lang=EN-GB><span style='font:7.0pt "Times New Roman"'>&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;

</span>i.<span style='font:7.0pt "Times New Roman"'>&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;

</span></span><span lang=EN-GB>level 3</span></p>

##### Other files

* MS-tutorial--Deep Learning for Signal and Information Processing.docx

<p class=MsoBodyText style='margin-left:.5in;text-indent:-.25in'>1)<span

style='font:7.0pt "Times New Roman"'>&nbsp;&nbsp;&nbsp;&nbsp;&nbsp; </span><b>Generative

deep architectures</b>, which are intended to capture high-order correlation of ... </p>

## Invisible paragraph

### Bug in Word HTML

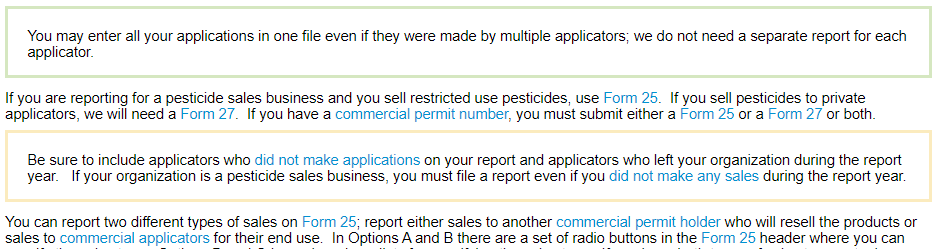
* Bug: paragraph text can be formatted as 'color.white', but the background is white, so the text is invisible
  + I've only seen this bug in a paragraph immediately after a table
* From Python Mammoth test files
  + test-Word-files\python-mammoth-Word-files\tables.docx

<p class=MsoNormal><span lang=EN-GB style='color:white'>Below</span></p>

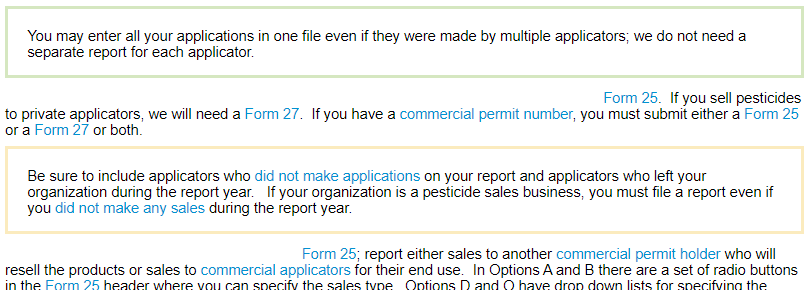
* I reproduced the bug in a Word file
  + table--testing-for-omitted-paragraph-after-table--v1.docx
  + There's no other color specifications for the HTML

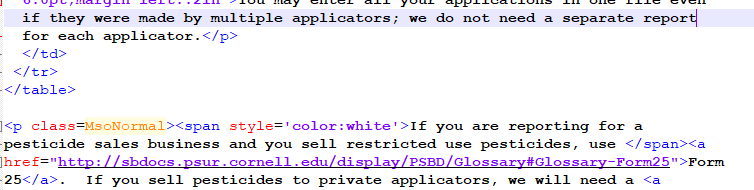
<p class=MsoNormal><span style='color:white'>paragraph 2</span></p>

* Found in
  + pesticide-report.docx
    - <http://sbdocs.psur.cornell.edu/download/attachments/4784135/OptionsDAndOUserGuide.docx?version=1&modificationDate=1449157135050&api=v2>
  + Word:



* + Word HTML





* <https://hogwartslive.com/privacy.html>
  + color:white is benign, since it's just for a space
  + after a table

</table>

</div>

<p class=MsoNormal style='text-align:justify;line-height:115%'><span

lang=en-US style='font-size:10.0pt;line-height:115%;font-family:"Arial",sans-serif;

color:white'>&nbsp;</span></p>

* <http://nigeria-law.org/>
  + color:white is benign, since it's just for a space
  + after a table

</table>

<p align=center style='margin:0cm;text-align:center'><span style='color:white;

mso-color-alt:windowtext'>&nbsp;</span></p>

* <http://jurnal.org/articles/2011/mat7.html>
  + Word HTML, filtered, but other divs were added, e.g., for ads
  + Minor problem, just for a "\_"

<html>

<head>

<title>Методы конечных разностей и конечных элементов в задачах электромагнитной совместимости</title>

<noindex>

<meta http-equiv=Content-Type content="text/html; charset=windows-1251">

<meta name=Generator content="Microsoft Word 14 (filtered)">

<style>

<p class=MsoNormal style='margin-left:0cm;text-align:justify;text-indent:35.45pt'>10.<span

style='font:7.0pt "Times New Roman"'>&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;

</span>Berenger J.-P. A perfectly matched layer for the absorption of

electromagnetic waves // J.Comput.Phys. – 1994. – Vol. 114, № 2. – P.<span

style='color:white'>\_</span>185-200</p>

* <http://webstratus.com/>
  + Word HTML, but not filtered
  + Minor problem, only affects a "."
  + After a <img tag

<img

border=0 width=431 height=106 src="default\_files/image001.png"

alt="Stratus-Header\_Traffic" v:shapes="Picture\_x0020\_1"></span><![endif]></span></a></p>

<p class=MsoNormal align=center style='text-align:center'><span

style='color:black;mso-themecolor:text1'>For questions or issues with Stratus <b

style='mso-bidi-font-weight:normal'>Traffic <span class=GramE>And</span>

Billing</b>, please email: <a href="mailto:trafficarhelp@cumulus.com"><span

style='color:black;mso-themecolor:text1'>trafficarhelp@cumulus.com</span></a></span><span

style='color:white;mso-themecolor:background1'>.<o:p></o:p></span></p>

* <http://www.ellington.k12.mo.us/ALUMNI.HTM>
  + Word HTML, but not filtered
  + In a table cell <td

<td><![endif]>

<div v:shape="\_x0000\_s1027" style='padding:3.6pt 7.2pt 3.6pt 7.2pt'

class=shape>

<p class=MsoNormal><span style='color:white;mso-themecolor:background1'>$10.00

each<o:p></o:p></span></p>

</div>

### Investigation, for creating a fix

#### PublicWWW.com searches

* "color:white" <table>
  + Bug is often with a paragraph just after a table

#### HTML for setting the background color

* Legitimate uses of color:white
  + <p> uses color:white, but intended, as background color is not white
* <https://www.bible.ca/canon.htm>
  + Pure Word HTML (filtered), legit use of color:white

<meta name=Generator content="Microsoft Word 15 (filtered)">

* + Table cell:

<td width=623 valign=top style='width:467.5pt;border:solid windowtext 1.0pt;

background:black;padding:0in 5.4pt 0in 5.4pt'>

<p class=MsoNoSpacing align=center style='text-align:center'><b><span

style='font-size:18.0pt;color:white'>The Canon of the Bible</span> [html removed]</p>

</td>

* <http://mrcophth.com/>
  + Not pure Word HTML, apprently from MS FrontPage
  + color:white appears to be a bug

<body bgcolor=white lang=EN-US link="#3333FF" vlink="#3333FF" style='tab-interval:

.5in' alink="#CC33CC">

<div class=Section1>

<p class=MsoNormal><span class=GramE><span style='font-size:7.5pt;color:white'>z</span></span><span

style='font-size:7.5pt;color:white'> Singapore National Eye Centre (SNEC) <span

class=SpellE>Moorfields</span> Eye Hospital World Ophthalmology Congress</span>

<o:p></o:p></p>

* <http://wwlln.net/>
  + Full Word HTML (not filtered), legit use of color:white

<body bgcolor=black lang=EN-US link="#CC33CC" vlink="#CC33CC" style='tab-interval:

.5in' alink="#ff0000">

<p class=MsoNormal style='mso-margin-top-alt:auto;mso-margin-bottom-alt:auto'><span

style='color:white'>&nbsp; <o:p></o:p></span></p>

<td width="22%" style='width:22.0%;background:black;padding:0in 0in 0in 0in'>

* <http://sbsc.com.vn/ContactUs.aspx>
  + Not pure Word HTML
  + This HTML might not be generated by Word
  + color:white set in <a> tag

<td style='background-color: #2d7caf;color:#fff; padding: 5px 10px; border-top:solid 1px White;font-size:14px;line-height:18px;margin-bottom:14px;font-family:Arial,sans-serif'><a style='color:white;'

### Word color-related formatting

#### Info from MS's HTML specs

* Word Borders and Shading
  + Microsoft Word allows the application of border and shading properties to text, paragraphs, sections, tables, and table cells. The following CSS style attributes correspond to Word formatting elements. Attribute Value in Word
    - **background-color :** The shading fill behind the text or art.
    - **background :** The shading fill of the object.
  + To preserve these effects, Word uses the following HTML style attributes. Note that each attribute has multiple values; for example, border-top uses a string to define the values (in order) of width, style, and color. Style Word property
    - **background** : Fill color of the element.
  + **color** : CSS Text See the CSS Level 2 Recommendation
  + **mso-background** : Office only Cell Formatting auto,<color>,windowtext
  + **mso-background-source** : Office only Cell Formatting auto
* Microsoft Word, Microsoft Excel, and Microsoft PowerPoint allow saving the background color or image to HTML.
  + Word
    - If the background is a color, Word implements the **bgcolor** attribute of the Body element using standard HTML colors

## Word HTML structure

### Overall

* The overall Word HTML structure is described in the Word HTML specs:
  + *Office HTML and XML File Formats*
    - *When a Microsoft Office document is saved as a Web page, a main HTML file and a number of related files are created.*
  + *Page Layout and Section Breaks*
    - *Many important Microsoft Word page layout settings are stored on a section-by-section basis within a document.*
* There can be mulitple div sections
  + From the Word HTML specs

<head><style> <!--

@page { document-level settings }

@page Section1 { first section settings }

div.Section1 { page: Section1; }

@page Section2 { second section settings }

div.Section2 { page: Section2; }

@page Section3 { third section settings }

div.Section3 { page: Section3; }

--></style></head>

<body>

<div class=Section1>first section data goes here</div>

<div class=Section2 >second section data goes here</div>

<div class=Section3> <third section data goes here</div>

</body>

### Top-most HTML tags, including <meta> tags

* table--testing-for-omitted-paragraph-after-table--v1.html
  + From my Word doc

<html>

<head>

<meta http-equiv=Content-Type content="text/html; charset=windows-1252">

<meta name=Generator content="Microsoft Word 15 (filtered)">

<style>

* charset=
  + charset=utf-8
  + charset=windows-1252
* test-Word-files\python-mammoth-Word-files--Word-HTML\comments.html
  + Word file has review comments
  + Word added a <script... element in <head>

<html>

<head>

<meta http-equiv=Content-Type content="text/html; charset=windows-1252">

<meta name=Generator content="Microsoft Word 15 (filtered)">

<style id="dynCom" type="text/css"><!-- --></style>

<script language="JavaScript"><!--

function msoCommentShow(anchor\_id, com\_id)

{

if(msoBrowserCheck())

* <http://dandanplay.com/bsintro.htm>

<html>

<head>

<meta http-equiv=Content-Type content="text/html; charset=utf-8">

<meta name=Generator content="Microsoft Word 15 (filtered)">

<style>

* <http://izapya.com/policy_en.html>

<html>

<head>

<meta http-equiv=Content-Type content="text/html; charset=utf-8">

<meta name=Generator content="Microsoft Word 15 (filtered)">

<title>Zapya Privacy Policy</title>

<style>

<!--

/\* Font Definitions \*/

[html removed]

-->

</style>

</head>

<body lang=ZH-CN link="#0563C1" vlink="#954F72" style="padding:1rem;">

<div class=WordSection1>

* <http://izapya.com/v3/about_us.html>
  + viewport does not appear to be generated by MS Word
  + It's common use seems to be in mobile-device support (small screen size) and with HTML email

<html>

<head>

<meta http-equiv=Content-Type content="text/html; charset=utf-8">

<meta name=Generator content="Microsoft Word 15 (filtered)">

<meta name="viewport" content="width=device-width,initial-scale=1.0,maximum-scale=1.0,user-scalable=0">

<title>About Us</title>

<style>

<!--

/\* Font Definitions \*/

* <http://nastooh.ir/docs/portfolio/newsagency-analytics-report.htm>
  + Arabic characters in title

<html>

<head>

<meta http-equiv=Content-Type content="text/html; charset=UTF-8">

<meta name=Generator content="Microsoft Word 15 (filtered)">

<title>&#1576;&#1575;&#1586;&#1578;&#1575;&#1576;

&#1605;&#1575;&#1607;&#1740;&#1575;&#1606;&#1607;

&#1582;&#1576;&#1585;&#1711;&#1586;&#1575;&#1585;&#1740;&#8204;&#1607;&#1575;</title>

<style>

<!--

/\* Font Definitions \*/

* test-Word-files\other-Word-files--Word-HTML\example-book-formatting-v2.html

<html>

<head>

<meta http-equiv=Content-Type content="text/html; charset=windows-1252">

<meta name=Generator content="Microsoft Word 15 (filtered)">

<title>Emma</title>

<style>

### div's

* Word features that create extra div's
  + footnotes, byte-order-mark, review comments
* test-Word-files\python-mammoth-Word-files--Word-HTML\footnote-hyperlink.html
  + Footnotes result in two additional div's, one nested in the other

<div class=WordSection1>

<p class=MsoNormal><a href="#\_ftn1" name="\_ftnref1" title=""><span

class=MsoFootnoteReference><span lang=EN-GB><span class=MsoFootnoteReference><span

lang=EN-GB style='font-size:11.0pt;line-height:115%;font-family:"Calibri",sans-serif'>[1]</span></span></span></span></a></p>

</div>

<div><br clear=all>

<hr align=left size=1 width="33%">

<div id=ftn1>

<p class=MsoFootnoteText><a href="#\_ftnref1" name="\_ftn1" title=""><span

class=MsoFootnoteReference><span lang=EN-GB><span class=MsoFootnoteReference><span

lang=EN-GB style='font-size:10.0pt;line-height:115%;font-family:"Calibri",sans-serif'>[1]</span></span></span></span></a><span

lang=EN-GB> <a href="http://www.example.com">Example</a></span></p>

</div>

</div>

* Byte order mark results in an additional div
  + test-Word-files\python-mammoth-Word-files--Word-HTML\utf8-bom.html
  + XML byte order mark - Google Search
  + Byte order mark - Wikipedia
    - <https://en.wikipedia.org/wiki/Byte_order_mark>

<div class=WordSection1>

<div style='border:none black 1.0pt;padding:0in 0in 0in 0in'>

<p class=MsoNormal>This XML has a byte order mark.</p>

</div>

</div>

## Unconventional characters

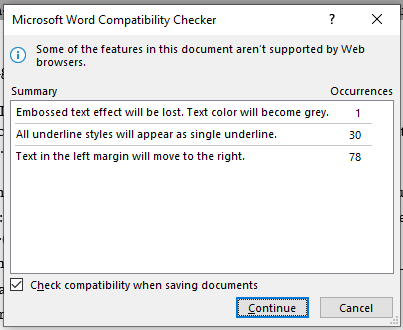
* How can I clean extra code out of Word HTML
  + <https://www.moorecreative.com/Articles/Detail/tabid/522/ArticleId/18/How-can-i-clean-extra-code-out-of-Word-HTML.aspx>
  + This app is one of our favorites because it also converts all non standard characters (like curly quotes, em and en dashes, Macintosh character issues, etc) into the proper ASCII.
* Smart quotes
  + Accessibility at Penn State | Cautions on Converting Word to HTML
    - <https://accessibility.psu.edu/microsoftoffice/microsoftword/wordhtml/>
    - If Smart Quotes are turned on, then they will be converted to a Unicode numeric character or left intact. Older browsers and screen readers may not be able to decipher these curly symbols. This issue also affects apostrophes and lengthened hypens.

# Shortcomings and bugs in Word's HTML

Some of these problems are described in the WWN User’s Guide

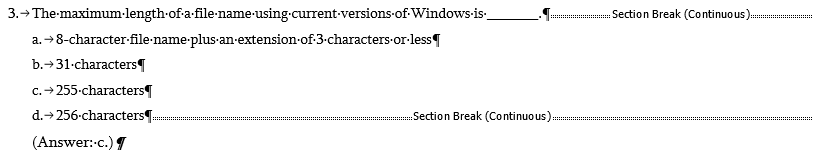
## Word messages about HTML formatting problems

* This GUI was displayed when saving .doc file as HTML
* computer-concepts-instructors-manual.docx
  + <http://virgil.azwestern.edu/~cvb/CIS120/Book%20Notes/Chapter.04.docx>



## Problems form spurious section-breaks in Word doc

* computer-concepts-instructors-manual.docx
  + <http://virgil.azwestern.edu/~cvb/CIS120/Book%20Notes/Chapter.04.docx>
* Suprious section breaks caused at least two problems in the Word HTML
  + Removing the section breaks fixed the problem in Word HTML



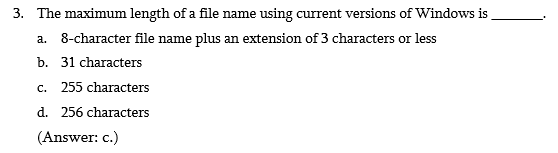
* The section breaks generated this Word HTML:

<span style='font-size:11.0pt;font-family:"Sylfaen",serif'><br clear=all

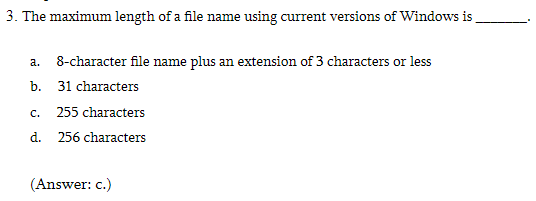
style='page-break-before:auto'>

</span>

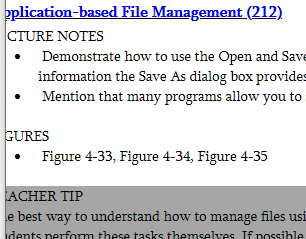
* Lines added before and after list
  + Word



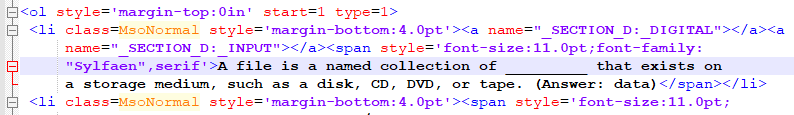
* + Word HTML



* Also causes left margin to be wrong for whole HTML doc

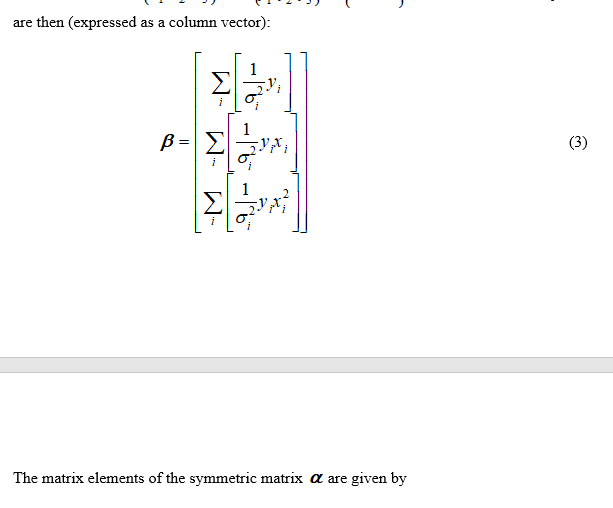


## Note: Word HTML uses <ol> and <li>

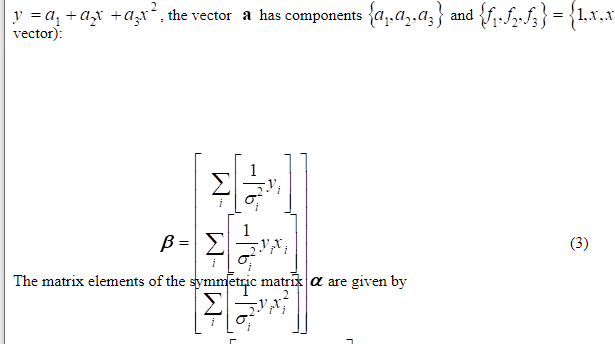


## Equation page-layout bug

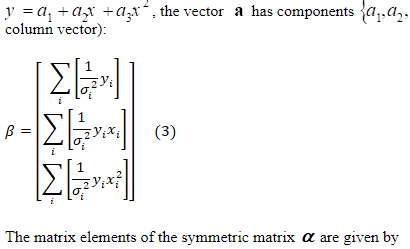
* Summary
  + When displayed in HTML, one equation is misplaced on the page
    - Word doc used old Word equation editor
    - The "(3)" for the equation was additional text
  + Fix
    - I used Word's feature to upgrade the equation for use with current Word equation editor
      * Right click on equation
    - I edited the equation to put the (3) in with the equation.
    - See: physics-tutorial--fixed.docx
* physics-tutorial.docx
  + <https://www.niu.edu/brown/_pdf/physics374_spring2021/l4-1-21.docx>
* Word



* Word HTML:

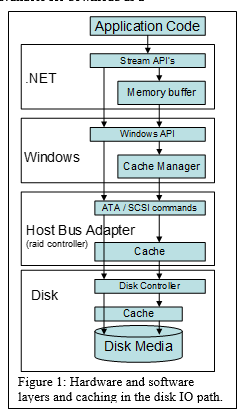


* Fixed, Word HTML

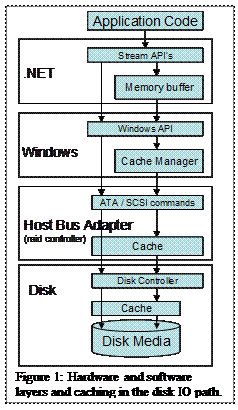


## Images can be become blurred

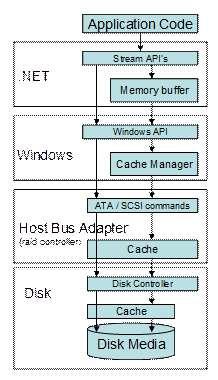
* Summary
  + Source image is actually a text box, with a picture and Figure caption as text
    - Problem: text boxes are converted to graphics images which are blurred
    - Solution:
      * Should be able to use a table instead of text box, to fix blurred Figure caption?
      * Image can be moved outside of the text box
  + Image blurred when document converted to Word HTML
    - Problem:
      * Embedded image in document is png
      * When doc is converted to HTML, image is converted to gif
      * Apparently gif conversion is blurry
    - Solution
      * Convert png to jpg, and embed the jpg image instead
      * When converting doc to HTML, Word will save the file as a jpg, and it displays legibly
  + Note: file is ".doc"
* Test file:
  + MS-tech-report--Sequential File Programming Patterns.doc
  + MS-tech-report--Sequential File Programming Patterns.htm
* In Word:
  + This is actually a text box with a png graphic image inside of it



* In Word HTML, in browser
  + Picture is now a linked gif file



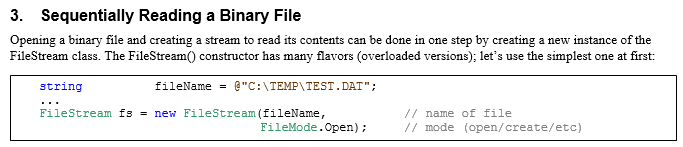
* Fix attempt: replace text box with just the image
  + Now, left-side fonts don't display well.



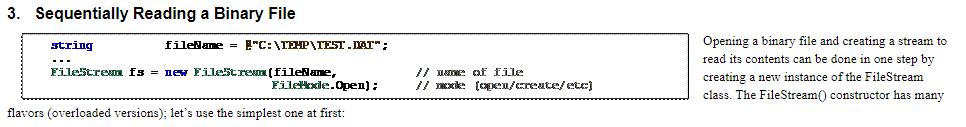
* MS-tutorial--Deep Learning for Signal and Information Processing.docx
  + The web-page images are much clearer here
  + Image techniques used:
    - Use a link to get image (e.g., gif ), from an external file
    - Embedded image is jpg

## Text boxes converted to illegible images, and layout problems

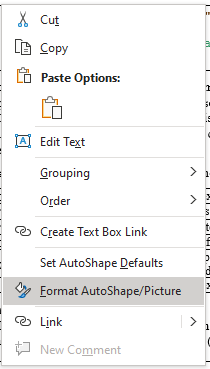
* Summary
  + legibility
    - An easy fix: use a table instead, with one cell
      * This was done later in the document, though with 2 cells, and it worked
    - Changing font-size in text box did not fix legibility
  + layout
    - Fixed by changing the text box layout specs
  + Note: file is ".doc"
* Test file:
  + MS-tech-report--Sequential File Programming Patterns.doc
  + MS-tech-report--Sequential File Programming Patterns.htm
* In Word

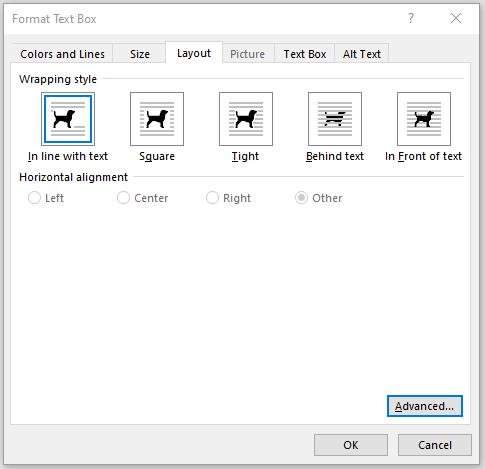


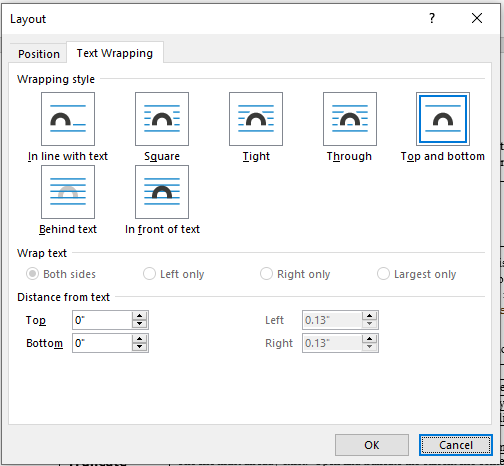
* In Word HTML, in browser



* Fix layout
  + Changed position relative to text. I was incorrect in Word doc, but didn't show-up there.

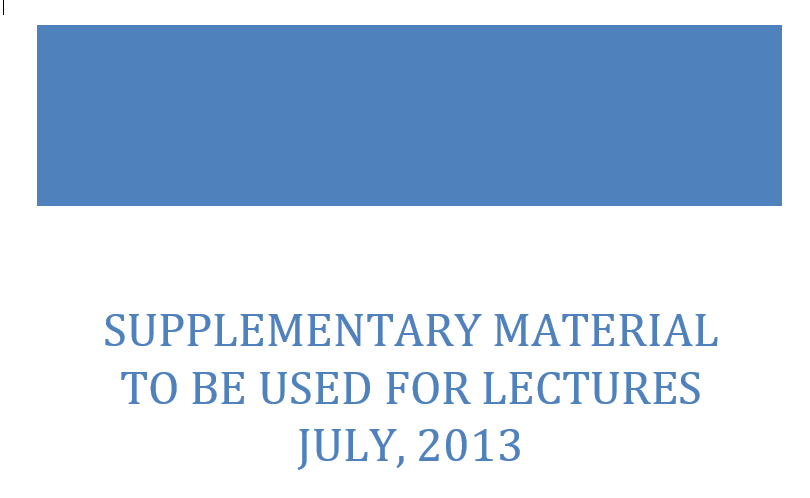






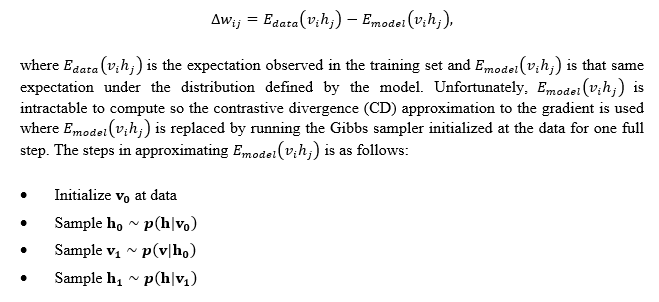
## Text-box not included in Word HTML document

* MS-tutorial--Deep Learning for Signal and Information Processing.docx
  + This figure is at the end of the document, and appears to be constructed of text-boxes within a text-box
  + It is not in the Word HTML

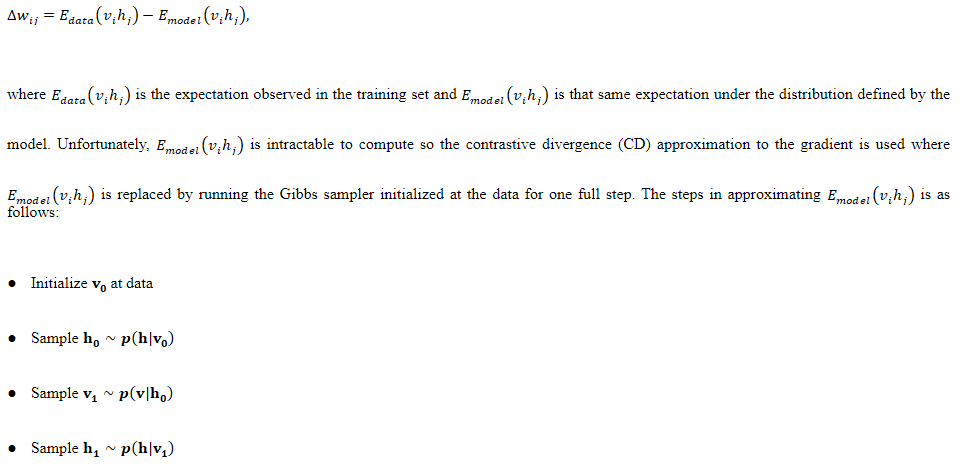


## Excessive space between lines

* MS-tutorial--Deep Learning for Signal and Information Processing.docx
* Word doc



* Word HTML



## References to page-numbers

* In the text, references to page-numbers are not meaningful in HTML
  + MS-tech-report--Sequential File--original.doc



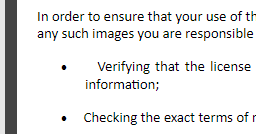
## Things HTML can't do, but Word can

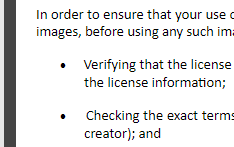
* HTML was orginally not designed to be a layout preserving format. If you want to offer a newsletter which was created using Word on a web page, best option may be to create a PDF from that and allow your audience to download it.
  + <https://stackoverflow.com/questions/8104230/converting-word-newsletter-to-html?rq=1>

# Word HTML: not a bug

## Effects of "justified" text

* For text that is "justified", the page width can cause beginning of line to be indented an extra amount
  + For "Justified" text, for each line in a paragraph (except last), the last letter is on the right margin.
  + Changing the page width via the slider can change that indentation





* MS-tutorial--Deep Learning for Signal and Information Processing.docx

