DOCX 2 TEI

Instructions and Example Document

**Tomaž Erjavec, Andrej Pančur**

2014-02-05

# Introduction

This document is meant as an exemplar and test Word file for a docx2tei profile of the [TEI Stylesheets](http://www.tei-c.org/Tools/Stylesheets/). It also functions as the source for a Word template (.dotx) than can serve for authoring new or editing existing Word documents (primarily books) with the intention of converting them to TEI. How Word structures are converted to TEI is here explained only briefly; to see the details it is best to compare the Word document with the generated TEI one.

This file and the associated profile, as well as a mini Web converter are available at <http://nl.ijs.si/tei/convert/>

In this document we give as examples the actual Word styles used, and when we refer to them, they are set in *italic*, e.g. the style *Quote*. To give TEI structures that these styles are converted to we use XPath and underline them, for <note place="left"> we write note[@place = "left"].

To get the best TEI results from a Word document using this profile:

* Very fancy formatting and esp. layout are not preserved in the TEI.
* The conversion heavily depends on using Word styles, so using “raw” formatting will not work well. The styles are here visually distinguished in a certain way, but changing their appearance to better suit individual use does not affect the conversion. Styles that are not used can be also deleted, to have less clutter in the Style gallery. Please note that English style names should be used; the conversion does currently not support styles names in other languages.
* For repetitive tasks (e.g. marking various types of names) it is convenient to define keyboard shortcuts; if only one style (e.g. *tei:name*) is being repeatedly applied, then CTRL-Y works as well.
* The text before the first heading goes to the element <front>, with styles Title, Subtitle, Author and Date being mapped to appropriate TEI elements. The rest of the text in this division ignored; the first heading and the rest becomes <body>.
* Some styles get extra processing, e.g. a series of paragraphs of style *Bibliography* are converted into a listBibl element containing a series of bibl elements.
* Styles starting with *tei:* are substituted by their TEI element, as explained in the Section on TEI elements.
* Some *tei:* styles have shorthand notation, e.g*. tei:orgName* goes to place[@type=’org’].

# Using standard Word formatting

This section reviews what kind of formatting we can do in standard Word to get appropriate TEI elements. The following section explains basic Word formatting (paragraphs, links, text effects) while the next two deal with character level and paragraph level styles. It is important to understand the distinction between the two, because the conversion to TEI is defined in terms of these two levels of styles. At the same time, Word does magic and can change one type of style to the other, which can lead to bad conversion results. When the TEI elements are not as expected it often helps to show Word formatting, i.e. pressing the “Show/Hide ¶” button and the Style gallery.

## Basic formatting

Plain paragraphs are converted to p. Empty paragraphs are removed even if they contain white-space, e.g:

All types of links (to [web pages](http://nl.ijs.si/tei/convert/), [mail addresses](mailto:tomaz.erjavec@ijs.si), and document-internal cross-references, e.g. to Section on Bugs) should be converted correctly. References to documents on disk will of course not work.

Formatting is converted to the value of hi/@rend: **bold**, *italic*, underline, and ~~strikethrough~~ are preserved, also if more than one style is used, e.g. ***italic bold underline***. Colours are also converted: rumeno ozadje, rdeča, svetlozelena, temno rdeča, oranžna, rumena, svetlomodra, modra, vijolična, ***podčrtana, krepka in nenavadna***. The exact details, such as the colour of underline and more **fancy text effects** are not preserved.

## Character level styles

Inside paragraphs we can have dates (with the *Date* style), which are converted to date, e.g. “It was a bright cold day in April, and the clocks were striking thirteen.”

## Paragraph level styles

Bulleted and numbered lists are supported, although the numbering style will not survive the conversion, e.g.:

1. First item
   1. Subitem
      1. Subsubitem

By using the paragraph level *Quote* style a quote can be produced:

My fake plants died because I did not pretend to water them.

Except for paragraph itself, there are no other general Word paragraph styles supported in the conversion.

## Notes

Margin Note Left

Margin Note Right

Here is a standard footnote[[1]](#footnote-1) and another,[[2]](#footnote-2) which should be converted without problems. We can also use endnotes[[3]](#endnote-1) although the difference between the two is moot in on-line editions.

Margin Note Inner

Margin Note Outer

Critical editions can also use marginal notes. Following one of the existing TEI profiles we define 4 *MarginNote* styles (*MarginNoteLeft, MarginNoteRight, MarginInner, MarginNoteOuter*), which are used to the left and right of this text. They are converted to note[@place = "margin\_xx"] where xx ∈ {left, right, inner, outer}. Note that the exact positioning of these notes is rather tricky.

## Figures

Figures and esp. tables are the more problematic aspects of conversion, as there are many ways to include them into a Word document. The pictures have to be embedded in the Word document. Because the conversion takes as input a Word document, references to external images are not supported.

The included pictures should be in as high resolution as possible – it is not a good idea to copy & paste them into Word, as this often loses resolution. Also, avoid embedding TIFF images if the TEI is to be afterwards converted to HTML as most Web browsers do not display TIFF.

If the figure has a caption, it should be made with “Insert Caption” so that it is in the correct style (*Caption*) and has automatic figure numbering. Note that the captions has to be *below* the image in order to get converted.

So, in short, the conversion supports embedded images with captions and references to them, c.f. Figure 1, which can be also referred to as the Figure below.

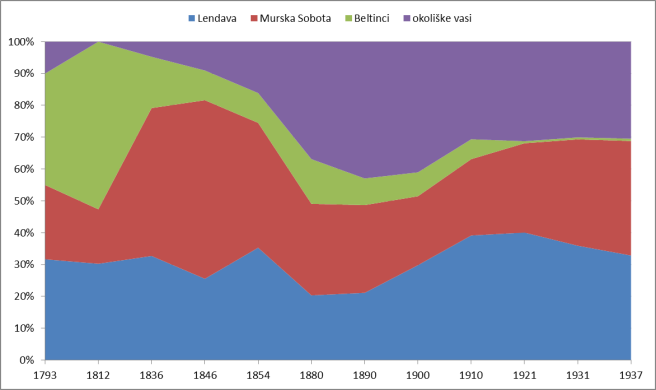


Figure 1 Some Statistics, as a picture

It is possible to have two images in one figure (i.e. with one *Caption*). As shown in Figure 2 they can be one above the other (i.e. they are separated by a paragraph mark) or they can be side by side, c.f. Figure 3.





Figure 2 Two images, one above the other



Figure 3 Two images side by side

Figures in Word can also be embedded Excel graphs, as is the case with Figure 4. However, this conversion currently does *not* work.

Figure 4 Embedded Excel example

We can also have pictures without captions. These are with this profile wrapped in figure, c.f. below:



## Tables

Tables, even somewhat more complicated ones (e.g. Table 2) can also be converted to TEI. However, the details of their layout and formatting will not be preserved. As with Figures, it is current not possible to convert embedded Excel spreadsheets.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Lendava** | **Murska Sobota** | **Beltinci** | **okoliške vasi** | **skupaj** |
| 1778 | *14* | *0* | *0* | *0* | *14* |
| 1793 | *19* | *14* | *21* | *6* | *60* |
| 1812 | *23* | *13* | *40* | *0* | *76* |

Table 1 A Simple table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **SLOVENSKE DEŽELE** | | | | |
|  | 144 | 192 | 182 |
|  | 96 | 89 | 145 |
| *Ljubljana* | *(74)* | *(76)* | *(95)* |
| **Sodni okraji z ozemlja današnje slovenske Primorske** | | | | |

Table 2 A more complex table with multi columns

## Indexes

Word also support the making of indexes and they are preserved in the conversion, as the example below shows (click on “Reveal formatting”, i.e. “Show ¶” to see the index marks):

“Here we are indexing the Web, Web services, and Web apps, but also bugs and errors. Note that the index terms can be in Word also formatted, which is lost in the TEI. We can have ranges though, like this.”

## Bibliography

Support for bibliography is quite basic – use the *Bibliography* style, as below, to get a listBibl element with nested bibl elements, i.e. listBibl/bibl+.

1. AUMÜLLER, Jutta: Assimilation: Kontroversen um ein migrationspolitisches Konzept. Bielefeld: Transcript Verlag, 2009.
2. BAGARIĆ, Marija: Obitelj Kapetanović. V: Pravednici među narodima [online], <http://www.geoskola.hr/hr/projekti/pravednici/kapetanovici.htm> (Accessed 27. 9. 2007).
3. BELLER, Steven: Wien und die Juden 1867-1938. Wien, Köln in Weimar: Böhlau, 1993.

## Page and line breaks

Page breaks are preserved in TEI, even soft ones. However, page breaks can be problematic, as they can appear inside any (even otherwise empty) element, like p, head, div. Hard line breaks also work, and are converted lb.  
There was a hard line break just before this sentence, and a hard page break follows it.

# TEI element styles

In addition to standard word styles, there is a special group of styles that start with “tei:” followed by (typically) the name of a TEI element. These styles are in the Word document given lots of eye watering effects to distinguish them from other text.

It some cases the styles are mapped to more complicated structures. An example is the *tei:lg* style: if a series of paragraphs use this style then each series ending with an empty paragraph is converted to lg, with l for the individual paragraph. To see the details for each style it is easiest to compare this file with the derived TEI.

## Paragraph level styles

These styles are paragraph level, i.e. they should mark complete paragraphs.

A citation is styled with *tei:cit* and can (as all other paragraph level styles) have included char level style, in this case *tei:bibl* to mark a bibliographic item:

A spectre is haunting Europe; the spectre of Communism. The Communist Manifesto (1848), by Karl Marx and Friedrich Engels

For poetry the *tei:lg* style should be used. This is converted to lg for each stanza, and l for a line in a stanza. Note that a series of stanzas can be styled with *tei:lg*, and an empty paragraph will separate the stanzas, as in the example below:

There once was a man from Nantucket

Who kept all his cash in a bucket.

But his daughter, named Nan,

Ran away with a man

And as for the bucket, Nantucket.

We can also have individual lines of poetry, without line group; for these the *tei:l* style should be used, e.g.:

There once was a man from Nantucket

Who kept all his cash in a bucket.

The *tei:sp* style should be used to mark a drama speech (sp). We implement the convention that the first paragraph goes to speaker and the rest are lines. As with a bibliography list, an empty paragraph will separate two speeches. For example:

Polonius

Though this be madness, yet there is method in’t.

Will you walk out of the air, my lord?

Hamlet

Into my grave.

## Character level styles

Character level Word styles map to various TEI phrase level elements:

* Basic style (elements) supported are *tei:bibl* (bibl), *tei:foreign* (foreign), and *tei:q* (q), e.g.:  
  “text Bibl text Foreign text “Quoted” text”.
* Named entities are marked up with *tei:name* (name), *tei:orgName* (name[@type="org"]), *tei:persName* (name[@type="org"]), *tei:orgName* (name[@type="org"]), e.g.:  
  “text Name text Person text Place text Organisation text”.
* Text-critical elements with eponymous names (note that the text marked with *tei:gap* goes to gap/desc, i.e. a description of the gapped text), e.g.:  
  “text Damage text Gap (description of) text Supplied text Surplus text Unclear text”.
* Note also the “Catch-word” at the bottom of this page, written with *tei:fwCatch* converted to fw[@type="catch"].

Note that if we have several such elements in a row there should be at least one character in between marked with *Normal* style in order to separate them. For example, (Pančur, 2011, 45-47; Erjavec, 2012, 30) will produce one bibl, whereas (Pančur, 2011, 45-47; Erjavec, 2012, 30) will produce two; the difference is in the semicolon, it is marked with the *tei:bibl* style in the former and *Normal* in the latter.[[4]](#footnote-3)

### Janus elements

So called Janus (two-faced) elements are used mostly in text-critical editions and are special in that they can represent two alternative paths through the text of the document. When taken as alternative encodings they are wrapped in a [subst](http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-subst.html) or [choice](http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-choice.html) element. In particular, a contiguous series of del and add elements gets in the TEI the parent element subst , while ordered pairs of abbr followed by expan, orig followed by reg and sic followed by corr get a parent choice element.

Examples of use:

* For deletions and additions in the source text use *tei:del* (del) and *tei:add* (add):  
  “Delete text Delete Add, text Add Delete, text Delete Add Delete Add, text Add”.
* For abbreviations and their expansions use *tei:abbr* (abbr) and *tei:expan* (expan):  
  “Abbreviation text Abbreviation Expansion, text Expansion”.
* For original and regularised text use *tei:orig* (orig) and *tei:reg* (reg):  
  “Orig text Orig Regularised, text Regularised”.
* For abbreviations and their expansions use *tei:abbr* (abbr) and *tei:expan* (expan):  
  “Sic text Sic Corr Sic Corr Sic Corr, text Corr”.

## Defining your own

It is possible to define new *tei:* styles, which will also get converted to the TEI element that follows the *tei:* prefix. However, in this case it is necessary to give them some visual features to distinguish them from the surrounding text and other styles, and, as with the other style, to use them only in contexts where the TEI element is allowed.

Catch-word

# Conversion to HTML

Currently default teitohtml is used. The intention is to supply two specialisations of the tei.css which would:

1. style the supported elements so that they would look the same as the *tei:\** styles in Word
2. nicely style all the elements, esp. the text critical ones, for the reader

# Conclusions and further work

We’ve used approach to authoring TEI documents via Word for many years now, but so far the workflow was from RTF to TEI with home grown XSLT, c.f. <http://nl.ijs.si/e-zrc/rtf2tei/>. Now we’re finally switching to standard Stylesheets, and this document and the associated profile are the first attempt.

The plan is to:

1. Fixing bugs & adding features, c.f. below
2. Make authoring CSS
3. Start working on 2HTML, reporting problems upstream
4. Better converter? Install OxGarage?
5. Split current profile into two:
   1. for social studies (focus on tables, figures, indexes, names, potentially soft pbs)
   2. for humanities (facsimile, text-critical and other *tei:* styles).

Including facsimiles:

* extend the idea of tei: styles to section headings (pass 4?):   
  if it is “tei:facsimile” this is the facsimile.
* it can be empty (or maybe just a p) – then a facsimile with hard pb number of surfaces is created
* can have figures directly (with captions or not)
* or has a *tei:surface* which is a numbered list of surfaces
* each surface has a description and / or URI of image (can be local if plugged into downloaded conversion); how are they marked? (if it looks like an uri it is an uri?)

## Bugs and missing features

* Embedded Excel spreadsheets are not converted.
* Problems with (phrase level) elements containing initial or final space, which should go outside the element, e.g. “foo<term> bar </term>baz” → “foo <term>bar</term> baz”, otherwise some injudicious use of normalize-space quickly produces run-together text, e.g. “foo<term>bar</term>baz”. Example: “text Sic text”
* pb’s at start / end of elements should be promoted upwards as far as they go (and containing elements zapped if then empty) //fixed locally in pass4//

# Appendix 1. Auto-generated sections

Word can auto-generate various tables of contents. Not clear if this is worth including in the TEI document (except if pageification is kept), as they would probably be better automatically generated from a tei:divGen.

## Index

bugs

horrible bugs. *See* bugs

error, 5

**formatting**, 5

Range

Subrange, 7

Web, 5

Web app, 5

Web service, 5

## Table of Figures

[Figure 1 Some Statistics, as a picture 3](#_Toc378364446)

[Figure 2 Two images, one above the other 3](#_Toc378364447)

[Figure 3 Two images side by side 4](#_Toc378364448)

[Figure 4 Embedded Excel example 4](#_Toc378364449)

## Table of Tables

[Table 1 A Simple table 4](#_Toc378364439)

[Table 2 A more complex table with multi columns 4](#_Toc378364440)

**Table of Contents**

[Introduction 1](#_Toc379023616)

[Using standard Word formatting 2](#_Toc379023617)

[Basic formatting 2](#_Toc379023618)

[Character level styles 2](#_Toc379023619)

[Paragraph level styles 2](#_Toc379023620)

[Notes 2](#_Toc379023621)

[Figures 3](#_Toc379023622)

[Tables 4](#_Toc379023623)

[Indexes 5](#_Toc379023624)

[Bibliography 5](#_Toc379023625)

[Page and line breaks 5](#_Toc379023626)

[TEI element styles 6](#_Toc379023627)

[Paragraph level styles 6](#_Toc379023628)

[Character level styles 6](#_Toc379023629)

[Janus elements 7](#_Toc379023630)

[Defining your own 7](#_Toc379023631)

[Conversion to HTML 8](#_Toc379023632)

[Conclusions and further work 8](#_Toc379023633)

[Bugs and missing features 8](#_Toc379023634)

[Appendix 1. Auto-generated sections 9](#_Toc379023635)

[Index 9](#_Toc379023636)

[Table of Figures 9](#_Toc379023637)

[Table of Tables 9](#_Toc379023638)

1. Small footnote (we don’t want a paragraph inside it). [↑](#footnote-ref-1)
2. A footnote with two paragraphs. This is the first.

   And this is the second. [↑](#footnote-ref-2)
3. This is an example of an endnote. [↑](#endnote-ref-1)
4. It is also possible to have character-level styles in notes, e.g., AUMÜLLER, Jutta: Assimilation: Kontroversen um ein migrationspolitisches Konzept. Bielefeld: Transcript Verlag, 2009. [↑](#footnote-ref-3)