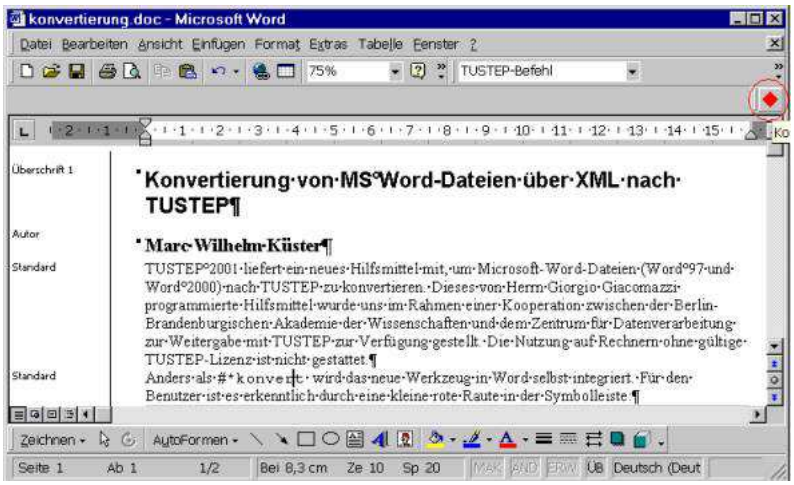
Konvertierung von WinWord-Dateien über XML nach TUSTEP

Zusammen mit der Version TUSTEP 2001 wird ein neues Hilfsmittel mitgeliefert, um Microsoft-Word-Dateien (Word 97 und Word 2000) nach TUSTEP zu konvertieren. Es wurde von Giorgio Giacomazzi 1996/97 zuerst für Word 6 und Word 95 entwickelt und im Rahmen einer Kooperation mit der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) auf Word 97 und Word 2000 portiert und zur Weitergabe mit TUSTEP zur Verfügung gestellt. Die Nutzung ist an eine gültige TUSTEP-Lizenz gekoppelt.

Das neue Werkzeug wird während der TUSTEP-Installation Word als Add-In hinzugefügt und ist also anders als `#*konvert` in Word selbst integriert. Technisch gesehen wird es in das »Start-up« Verzeichnis von MS Word integriert und so bei jedem Start des Textverarbeitungsprogramms automatisch gestartet. Für den Word-Benutzer ist es erkenntlich durch eine kleine rote Raute in der ein- und ausschaltbaren Symbolleiste »Konvertieren«:



Durch Anklicken dieser Raute wird die Word-Datei in eine wohlgeformte XML-Datei überführt, die dann mit `#umwandle` in TUSTEP importiert werden kann. Alternativ kann man den gleichen Effekt auch durch den neuen Eintrag »Speichern als XML« im Menupunkt »Datei« erreichen:

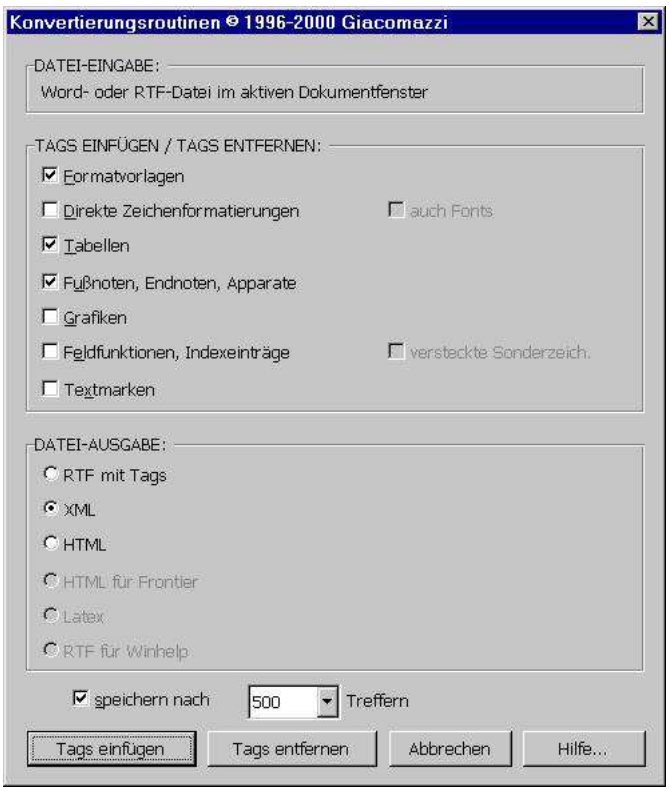
Die dadurch aktivierten Konvertierungsroutinen explizieren beliebige (auch benutzerdefinierte) Formatvorlagen, die in dem zu konvertierenden Word-Dokument benutzt werden.

Für diesen Artikel, dessen Ursprungsversion zur Illustration in Word geschrieben und dann konvertiert wurde, wurden zwei spezielle Formatvorlagen definiert, eine Absatzformatvorlage »Autor« und eine Zeichenformatvorlage »TUSTEP-Befehl«. Letztere kennzeichnet spezielle TUSTEP-Befehle wie z. B. `#*konvert` im Text. In der resultierenden XML-Datei erscheinen diese als

<TUSTEP-Befehl>#*konvert</TUSTEP-Befehl>

und können so sowohl im Satz typographisch geeignet ausgewertet werden als auch etwa für ein Register automatisch extrahiert werden.

Das Add-In bietet dem Nutzer folgende Optionen:



Im Normalfall wird man die Einträge so belassen, wie sie voreingestellt sind, und einfach »Tags einfügen« wählen. Möchte man zusätzlich andere Kategorien wie etwa Grafiken oder Feldfunktionen explizieren, so wählt man sie in der Maske aus. Direkte Zeichenformatierungen sollten nur aktiviert werden, wenn zusätzlich zu den Formatvorlagen auf Zeichenebene Formatierungen vorgenommen wurden oder vermutet werden (Diagnose), die für die Interpretation des Dokuments zentral sind.

So gut wie nie wird die Dateiausgabe auf etwas anderem als »XML« stehen. Die Optionen »rtf mit Tags« und »HTML« werden nur in Ausnahmefällen Verwendung finden.

Mit »Tags einfügen« und ggf. der Nachfrage, ob eine schon vorhandene Datei überschrieben werden darf, wird die eigentliche Konvertierung gestartet. Nach deren Ende fragt das Werkzeug noch einmal nach, ob die so entstandene Datei direkt als Textdatei mit XML-Auszeichnungen gespeichert werden soll (»Ja«) oder ob man vorzieht, sie sich zunächst einmal als Word-Datei mit bereits explizierten Formatvorlagen anzuschauen (»Nein«).

Wählt man letzteres, was vor allem für sukzessive Konvertierung (s. u.) sinnvoll ist, so muss man später selbst dafür sorgen, dass die Datei im reinen Textformat abgespeichert wird, kann dafür aber Textdateien mit nicht-lateinischen Schriftzeichen in Unicode abspeichern und dann mit `#umwandle, code=unicode` in TUSTEP einlesen. (Zur Prüfung des XML-Dokuments durch einen XML-Parser muß in diesem Fall das Atrribut encoding='ISO-8859-1' angepasst werden.)

Im Gegensatz zu dem ab Word 2000 vorhandenen Webexport unter Benutzung von Stylesheets gibt dieses Werkzeug direkten und selektiven Zugriff auf die Namen der Formatvorlagen und liefert unmittelbar wohlgeformtes XML.

Mit dem neuen Konvertierungswerkzeug lassen sich prinzipiell beliebige Word- (und rtf-)-Dateien nach XML überführen. Die Nutzbarkeit der erzeugten XML-Datei hängt dabei aber natürlich stark von der Qualität der verwendeten Formatvorlagen ab. Je eher diese den Ansprüchen sachlich orientierter Textauszeichnung entsprechen, desto einfacher wird später die Weiterverarbeitung sein.

Das Resultat sieht wie folgt aus:

```
<?xml version=»1.0« encoding='ISO-8859-1' standalone='yes'?>

<!-- Konvertierungsroutinen (c) 1996-2000 Giacomazzi -->
<!-- TableToHtml, Yura Lesiuk 1996 -->
<!-- EINGABE: konvertierung.doc, 19.01.01 11:47 -->
<!-- AUSGABE: konvertierung.xml, 19.01.01 11:48 -->
<!-- Einfügen von Tags für Fuß-/Endnoten, Tags für Tabellen, Tags für
Formatvorlagen, Tags für Grafiken, Tags für direkte Formatierungen,
-->
<!-- TAGS, DIE EINGEFÜGT WORDEN SIND: -->
<!-- (1) <DOKUMENT> -->
<!-- (2) <p> -->
<!-- (3) <Autor> -->
<!-- (4) <Überschrift_1> -->
<!-- (5) <TUSTEP-Befehl> -->
<!-- (6) <b> -->
<!-- EIGENSCHAFTEN DES STANDARD-ABSATZES: Times New Roman, 12 pt,
Deutsch (Deutschland), Linksbündig, Zeilenabstand einfach, Absatz-
kontrolle -->
<!-- EIGENSCHAFTEN ANDERER FORMATVORLAGEN -->
<!-- Autor: Überschrift 4 + Ebene 4 -->
<!-- Überschrift 1: Standard + Schriftart: Arial, 16 pt, Fett, Unter-
schneidung ab 16 pt, Abstand Vor 12 pt Nach 3 pt, Absätze nicht trennen,
Ebene 1 -->
<!-- TUSTEP-Befehl: Absatz-Standardschriftart + Schriftart: Courier
New -->

<!-- ENDE PROTOKOLL -->

<DOKUMENT>

<Überschrift_1>Konvertierung von MS Word-Dateien über XML nach TU-
STEP</Überschrift_1>
<Autor>Marc Wilhelm Küster</Autor>
<p>TUSTEP 2001 liefert ein neues Hilfsmittel [...] zur Verfügung
gestellt.</p>
<p>Anders als <TUSTEP-Befehl>#*konvert</TUSTEP-Befehl> wird das neue
Werkzeug in Word selbst integriert. Für den Benutzer ist es erkenntlich
durch eine kleine rote Raute in der Symbolleiste:</p>
[...]
</DOKUMENT>
```

Der Konverter greift im Prinzip auf alle Word-Formatierungen zu. Hingewiesen sei auf einige Besonderheiten:

– »Direkte Formatierungen«: das sind Zeichenformatierungen (Fett, Schriftart usw.), die in der zugrundeliegenden Formatvorlage nicht enthalten sind, sondern vom Benutzer einem markierten Textbereich direkt zugewiesen wurden.
– »Tabellen« werden in HTML-Tabellen konvertiert.
– Absätze mit der von Word automatisch zugewiesenen Formatvorlage »Standard« werden mit den Tags . . . ausgezeichnet.

Exkurs: Selektiv-sukzessive Konvertierung von Formatvorlagen

Bei komplexen bzw. umfangreichen Word-Dokumenten empfiehlt sich die sukzessive Konvertierung nach Word-Kategorien, also z. B. erst die Fußnoten, dann die Tabellen und letztendlich die Formatvorlagen. Der Konverter lässt zwar zu Testzwecken eine nachträgliche Konvertierung der Fußnoten zu. Diese sollten aber, wie voreingestellt, immer im ersten Durchgang konvertiert werden.

Standardmäßig werden alle im Dokument benutzten Formatvorlagen expliziert. Alternativ können, auch sukzessive, nur bestimmte Formatvorlagen konvertiert werden. Dies geschieht automatisch, wenn im Quellverzeichnis eine einfache Textdatei gleichen Namens, aber mit der Erweiterung ».FV« (Formatvorlagen) gefunden wird, welche die Namen der zu konvertierenden Dateien jeweils in einer eigenen Zeile enthält, z. B.:

Quelldatei: KONVERTIERUNG.DOC
FV-Datei: KONVERTIERUNG.FV

Wenn die FV-Datei z. B. die Zeile »TUSTEP-Befehl« enthielte, würde nur die Formatvorlage »TUSTEP-Befehl« behandelt. Zur sukzessiven Konvertierung darf die erstellte XML-Datei bei der abschließenden Rückfrage an den Benutzer nicht definitiv in XML gespeichert werden, da die Word-internen Formatierungen damit entfernt werden.

Marc Wilhelm Küster
marc.kuester@zdv.uni-tuebingen.de