

## Abstract

"Extension of Word 2000 for the implementation and analysis of **Unicode characters in Higher Planes** like Hieroglyphs and other Archaic characters;  
**Macro for Analysing** all kind of characters in a Word text with showing off the English character names according to ISO/IEC 10646".

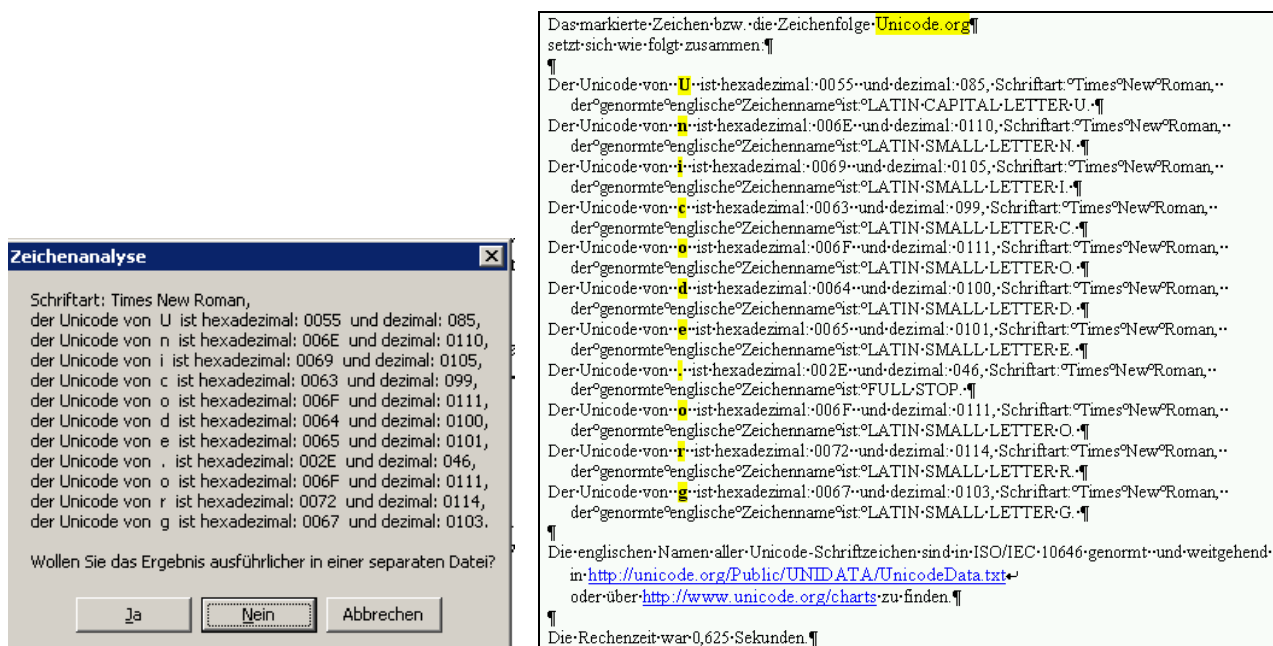
This book offers macros for the extension of Word 2000 to make it able to handle Unicodes of higher Unicode-planes like Hieroglyphs like  and  and Archaic characters.

For this Windows XP had to be adapted to cope with Surrogates since Windows XP works with UTF 16. Then I could search for fonts. There is no such thing as one single font for all characters but only for a rather small range (see chapter 4).

Now I had to enlarge Word 2000 by macros to write those characters of higher Unicode planes. I added the function of inserting Unicode characters by writing the hexcode and pressing Alt + C, which is normally not yet part of Word 2000, and the function of inserting characters by using the decimal code (see chapter 5).

Chapter 6 contains a **Macro for Analysing** characters in all versions of Word (Word 2000 to Word 2010) which analyses up to 44 characters in a Word document with one mouse click. It gives out the Unicode numbers (code point) of the marked characters and can also display their English names.

For example: When marking with the cursor "**Unicode.org**" and starting the macro (I have assigned to the macro the key-combination "Control + Shift + C)", I will get the codes (see Fig. 1, left hand side). In most cases this result is enough. When pressing "Ja" ('yes'), the macro looks up the English names of the characters in the Internet on the webpage of Unicode.org (see Fig. 1, right hand side):



Das markierte Zeichen bzw. die Zeichenfolge **Unicode.org** setzt sich wie folgt zusammen.

Der Unicode von **U** ist hexadezimal: 0055 und dezimal: 085, Schriftart: Times New Roman, der genormte englische Zeichenname ist: LATIN-CAPITAL-LETTER-U.

Der Unicode von **n** ist hexadezimal: 006E und dezimal: 0110, Schriftart: Times New Roman, der genormte englische Zeichenname ist: LATIN-SMALL-LETTER-N.

Der Unicode von **i** ist hexadezimal: 0069 und dezimal: 0105, Schriftart: Times New Roman, der genormte englische Zeichenname ist: LATIN-SMALL-LETTER-I.

Der Unicode von **c** ist hexadezimal: 0063 und dezimal: 099, Schriftart: Times New Roman, der genormte englische Zeichenname ist: LATIN-SMALL-LETTER-C.

Der Unicode von **o** ist hexadezimal: 006F und dezimal: 0111, Schriftart: Times New Roman, der genormte englische Zeichenname ist: LATIN-SMALL-LETTER-O.

Der Unicode von **.** ist hexadezimal: 002E und dezimal: 046, Schriftart: Times New Roman, der genormte englische Zeichenname ist: FULL-STOP.

Der Unicode von **o** ist hexadezimal: 006F und dezimal: 0111, Schriftart: Times New Roman, der genormte englische Zeichenname ist: LATIN-SMALL-LETTER-O.

Der Unicode von **r** ist hexadezimal: 0072 und dezimal: 0114, Schriftart: Times New Roman, der genormte englische Zeichenname ist: LATIN-SMALL-LETTER-R.

Der Unicode von **g** ist hexadezimal: 0067 und dezimal: 0103, Schriftart: Times New Roman, der genormte englische Zeichenname ist: LATIN-SMALL-LETTER-G.

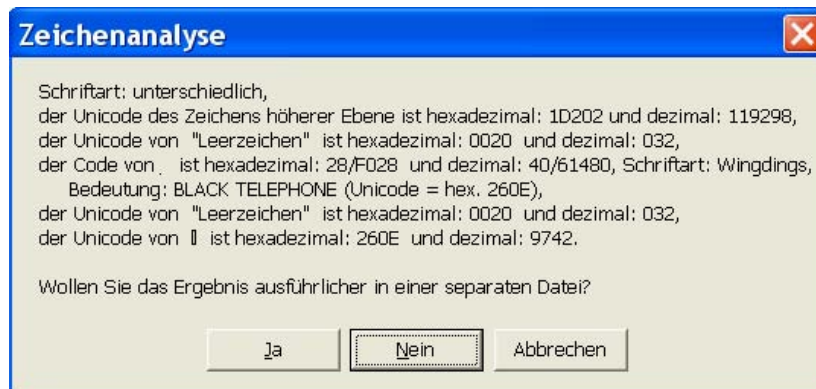
Die englischen Namen aller Unicode-Schriftzeichen sind in ISO/IEC 10646 genormt und weitgehend in <http://unicode.org/Public/UNIDATA/UnicodeData.txt> oder über <http://www.unicode.org/charts> zu finden.

Die Rechenzeit war 0,625 Sekunden.

Wollen Sie das Ergebnis ausführlicher in einer separaten Datei?

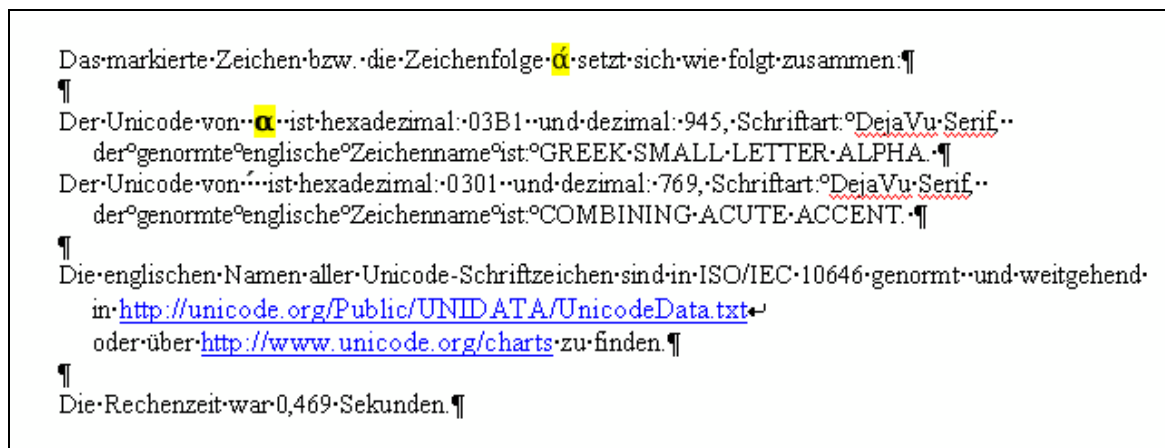
Fig. 1 Screenshot of the analysis

This also works for very unknown characters like: ✕ ☎ ☎ (see Fig. 2). The macro gives the also meaning of Symbol and Wingding characters. Perhaps here I should press "Ja" to get the name also of the first character of the marked sample<sup>1</sup> which belongs to a higher Unicode plane.



**Fig. 2: Screenshot for the analysis of unknown characters**

The analysing macro recognises characters with diacritical marks (see Fig. 3). The advantage in comparison to Word (even Word 2007) is that the macro always analyses the character and the combining accent together. There is no more danger of analysing only one part of the combined letter.



**Fig. 3: Screenshot of the analysis of a small letter Alpha with acute**

<sup>1</sup> GREEK VOCAL NOTATION SYMBOL-3

Fig. 4 shows the analysis of two Hieroglyphs (𐀀 𐀁) as an example of a higher Unicode-plane:

```

Das-markierte-Zeichen-bzw.-die-Zeichenfolge 𐀀·𐀁
setzt-sich-wie-folgt-zusammen.¶
¶
Der-Unicode-des-Zeichens-höherer-Ebene-𐀀-ist-hexadezimal:13000-und-dezimal:77824,
(der-Unicode-des-Surrogat-1-ist-hexadezimal:D80C-und-dezimal:55308,
der-Unicode-des-Surrogat-2-ist-hexadezimal:DC00-und-dezimal:56320),
der-genormte-englische-Zeichename-ist:EGYPTIAN-HIEROGLYPH-A001,
die-Schriftart-ist-Gardiner.¶
Der-Unicode-von-"Leerzeichen"-ist-hexadezimal:0020-und-dezimal:032,
Schriftart:TimesNewRoman,der-genormte-englische-Zeichename-ist:SPACE.¶
Der-Unicode-des-Zeichens-höherer-Ebene-𐀁-ist-hexadezimal:13050-und-dezimal:77904,
(der-Unicode-des-Surrogat-1-ist-hexadezimal:D80C-und-dezimal:55308,
der-Unicode-des-Surrogat-2-ist-hexadezimal:DC50-und-dezimal:56400),
der-genormte-englische-Zeichename-ist:EGYPTIAN-HIEROGLYPH-B001,
die-Schriftart-ist-Gardiner.¶
¶
Die-englischen-Namen-aller-Unicode-Schriftzeichen-sind-in-ISO/IEC-10646-genormt-und-weitgehend-
in-http://unicode.org/Public/UNIDATA/UnicodeData.txt
oder-über-http://www.unicode.org/charts-zu-finden.¶
¶
Die-Rechenzeit-war-0,719-Sekunden.¶

```

**Fig. 4: Two Hieroglyphs as a sample for a higher Unicode plane**

The macro can also handle combination from left to right like Latin characters and from right to left like Arabic and Hebrew characters (see Fig. 5):

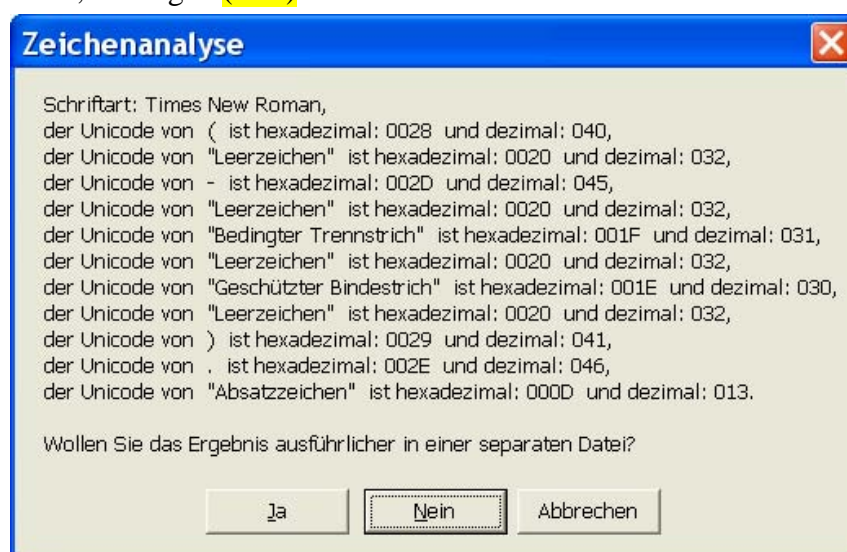
```

Das-markierte-Zeichen-bzw.-die-Zeichenfolge in: 𐤀-(HEBREW) setzt-sich-wie-folgt-zusammen.¶
¶
Der-Unicode-von-i-ist-hexadezimal:0069-und-dezimal:00105,Schriftart:TimesNewRoman,
der-genormte-englische-Zeichename-ist:LATIN-SMALL-LETTER-I.¶
Der-Unicode-von-n-ist-hexadezimal:006E-und-dezimal:00110,Schriftart:TimesNewRoman,
der-genormte-englische-Zeichename-ist:LATIN-SMALL-LETTER-N.¶
Der-Unicode-von-"Leerzeichen"-ist-hexadezimal:0020-und-dezimal:0032,Schriftart:TimesNew
Roman,der-genormte-englische-Zeichename-ist:SPACE.¶
Der-Unicode-von-𐤀-ist-hexadezimal:05D0-und-dezimal:1488,Schriftart:MicrosoftSansSerif,
der-genormte-englische-Zeichename-ist:HEBREW-LETTER-ALEF.¶
Der-Unicode-von-𐤁-ist-hexadezimal:05E8-und-dezimal:1464,Schriftart:MicrosoftSansSerif,
der-genormte-englische-Zeichename-ist:HEBREW-POINT-QAMATS.¶
Der-Unicode-von-"Leerzeichen"-ist-hexadezimal:0020-und-dezimal:0032,Schriftart:TimesNew
Roman,der-genormte-englische-Zeichename-ist:SPACE.¶
Der-Unicode-von-(ist-hexadezimal:0028-und-dezimal:0040,Schriftart:TimesNewRoman,
der-genormte-englische-Zeichename-ist:LEFT-PARENTHESIS.¶
Der-Unicode-von-H-ist-hexadezimal:0048-und-dezimal:0072,Schriftart:TimesNewRoman,
der-genormte-englische-Zeichename-ist:LATIN-CAPITAL-LETTER-H.¶
Der-Unicode-von-E-ist-hexadezimal:0045-und-dezimal:0069,Schriftart:TimesNewRoman,
der-genormte-englische-Zeichename-ist:LATIN-CAPITAL-LETTER-E.¶
Der-Unicode-von-B-ist-hexadezimal:0042-und-dezimal:0066,Schriftart:TimesNewRoman,
der-genormte-englische-Zeichename-ist:LATIN-CAPITAL-LETTER-B.¶
Der-Unicode-von-R-ist-hexadezimal:0052-und-dezimal:0082,Schriftart:TimesNewRoman,
der-genormte-englische-Zeichename-ist:LATIN-CAPITAL-LETTER-R.¶
Der-Unicode-von-E-ist-hexadezimal:0045-und-dezimal:0069,Schriftart:TimesNewRoman,
der-genormte-englische-Zeichename-ist:LATIN-CAPITAL-LETTER-E.¶
Der-Unicode-von-W-ist-hexadezimal:0057-und-dezimal:0087,Schriftart:TimesNewRoman,
der-genormte-englische-Zeichename-ist:LATIN-CAPITAL-LETTER-W.¶
¶
Die-englischen-Namen-aller-Unicode-Schriftzeichen-sind-in-ISO/IEC-10646-genormt-und-weitgehend-
in-http://unicode.org/Public/UNIDATA/UnicodeData.txt
oder-über-http://www.unicode.org/charts-zu-finden.¶
¶
Die-Rechenzeit-war-0,703-Sekunden.¶

```

**Fig. 5: A Hebrew letter with Latin letters in the same line**

Last not least the analysing macro also recognizes special marks (signs) like NoBreakSpace, NoBreakHyphen, Soft Hyphen, Tabulator, Paragraph, etc. even when they are only Control characters in Word, see Fig. 6 (- -).



**Fig. 6: Special marks**

The limitation of my analyzing macro for Word is, that it cannot give you the names of Chinese, Japanese or Korean syllable-characters.

The booklet can be downloaded:


- [1] Gast, Hanna-Chris: "Erweiterung von Word 2000 zur Darstellung und Analyse von Schriftzeichen höherer Unicode-Ebenen sowie Erstellung eines Makros für Word 2000 bis Word 2007 zur Analyse von Schrift- und Sonderzeichen mit Ausgabe des Schriftzeichen-Namens nach ISO/IEC 10646";  
Beilage zum Siebener-Kurier Nr. 60 (August 2010), ISSN 0948-6089;  
*In the Internet as download:*  
<http://www.siebener-kurier.de/chris-aufsaeetze/Word-Erweiterung-Unicode-Makros.pdf>.

*The macro for analysing characters of chapter 6 for Word 2000 up to Word 2010 can be separately downloaded under:*

[http://www.siebener-kurier.de/chris-aufsaeetze/Unicode\\_Analyse.txt](http://www.siebener-kurier.de/chris-aufsaeetze/Unicode_Analyse.txt);

**Unfortunately this book is only available in German language.**

Chris

P.S. If you cannot recognize a character in a website or in a pdf-document, just copy it into a word-file and analyse it with this macro. The macro also works well, if you only see "blank boxes" ("blank place-holders") like " " or .