The AGNT Project Report—Q2 2017

As a licensee or friend of AGNT or ANLEX, we would like to update you once a quarter about our continuing work to enhance and perfect these databases and about our plans for the future.

The Project. *The AGNT Project Report—Q3 2008* introduced the team, outlined ongoing tasks, and discussed potential tasks**.**

Unicode Composite Characters and Diphthongal Segments of Long Vowels

John Hughes and Timothy Friberg

This article explores the problem of creating properly formed composite characters in Unicode, focusing on diphthongal segments of long vowels (α or η or ω) in Greek.

A *composite character* is any character that consists of a consonant or vowel (one of the letters of an alphabet) + one or more diacritical marks, which could be above or below or—in the case of Hebrew—within the character. (Sometimes composite characters are referred to as *accented glyphs*; see the article at the first link below.)

Diphthongal segments of long vowels followed by iota (ι) in Greek, whether Classical, Koine or Byzantine, typically were written in either one of two ways. First, the long vowel was followed by an adscript ι, whether αι, ηι, ωι. Or it was written with the iota subscripted to the first vowel: ᾳ, ῃ, ῳ.

Instances of capitalization wherein the first segment of the diphthong used the corresponding uppercase character similarly had two possibilities of expression, analogously, Αι, Ηι, Ωι, or ᾼ, ῌ, ῼ. Some printed works in Greek use one system, some use the other.

In our ANLEX, it is important for us to display both systems and that side by side. For example, **ᾅδης**, **ου**, **ὁ** (also **ᾍδης**, **Ἅιδης)**, the keyword for the Greek behind Hades. We need to show for the reader that different versions of text treat the diphthongal phenomenon in either of two ways.

Now it turns out that this is not an editor problem, but more basically the Unicode font behind what one sees on the screen. Times New Roman, for example, allows for both expressions of the diphthong, whereas Galatia SIL and Gentium only allow for the uppercase adscript solution and disallow the uppercase subscript solution.

In preparing our AGNT materials, we can use a font, say Times New Roman, which allows both expressions. But we prefer Gentium or Galatia SIL as far more attractive fonts.

But it isn't enough that our hand is forced to embrace the ugly. It is further complicated in that we do not determine what fonts our vendors use. If we produce something in Times New Roman, which appears on our screens in both forms, but a vendor uses Galatia SIL, what he displays on the user's screen will be decidedly wrong.

So here are the two problems we face. First, not all fonts allow both expressions of writing of diphthongs. Second, AGNT vendors, whom we do not police with respect to fonts used to display our AGNT materials, may indeed choose a font that is resistant to "true" iota subscripting following uppercase alpha, eta, or omega. Additionally, some Bible software allows users to select any installed Unicode Greek font.

Unicode fonts (e.g., Gentium, Galatia SIL) are, essentially, databases. Computer software (e.g., Word, your e-mail program, your web browser, a Bible software program) are programmed—more or less successfully—to work with Unicode fonts. In other words, the fonts don't "do" anything; the programs do something with the fonts—render them more or less correctly. It is not difficult to specify in Unicode what elements should be combined to form a composite character. The "trick" is for the software (e.g., Word) in which a font that contains composite characters (e.g., Gentium, Galatia SIL) is used to *render* the composite characters correctly. In other words, it is not a font-coding issue but a software-rendering issue. But**—**and this is a big "but"—Unicode fonts may be improperly created; they may be incomplete, as we noted above regarding Gentium and Galatia SIL, which do not support the uppercase subscript solution for diphthongal segments of long vowels.

From a font-creation perspective, composite characters can be *precomposed* or *decomposed*. Some composite characters are *precomposed*, which means they are specified by a single Unicode number, e.g., an "e" with acute accent = U+00E9 (Unicode numbers are hex, not decimal). There are two reasons for using precomposed characters to create composite characters. (a) "To aid computer systems with *incomplete* Unicode support, where equivalent decomposed characters may render incorrectly" (emphasis mine; from the "Precomposed Character" article at the link below), and (b) to reduce the number of code points needed to represent all compound character possibilities in a given language (see "Ready-made Versus Composite Characters" at the link below).

Composite characters can be *decomposed* into their base letter, e.g., "e," + the diacritical, e.g., acute accent, = U+0065+U+0301. This means that composite characters can be created in Unicode by combining a base letter with one or more diacriticals. How this translates into typing strokes would be up to the software (keyboard program) that uses the font inside of the larger application (e.g., Word).

So, a given composite character (e.g., A + angstrom in Swedish) often can be specified in Unicode in more than one way: precomposed and decomposed (see "Precomposed Character" article, and, especially, "Unicode Equivalence" article at the links below).

"Some Unicode implementations still have difficulties with decomposed characters. In the worst case, combining diacritics may be disregarded or rendered as unrecognized characters after their base letters, as they are not included in all fonts. To overcome the problems, some applications may simply attempt to replace the decomposed characters with the equivalent precomposed characters" (see "Precomposed Character" article at the link below). One reason not to use highly decomposed character sets is that this "would introduce challenges for searching and editing software and require more bytes of encoding per document" (ibid). So an ideal Unicode Greek font would provide all the necessary composite characters as *precomposed* characters, representing each such precomposed character with a single hex number, a single "code point."

Essentially, the only "fix" for the problem described above is for the creators of Gentium and Galatia SIL to add precomposed subscript versions of Α + ι, Η + ι, and Ω + ι to their fonts. In other words, they should make their incomplete fonts more complete!

References:

- https://fontforge.github.io/accented.html. "Building Accented and other Composite Glyphs."
- https://en.wikipedia.org/wiki/Precomposed_character. "Precomposed Characters"
- https://en.wikipedia.org/wiki/Unicode#Ready-made_versus_composite_characters. "Ready-made Versus Composite Characters" in the larger article "Unicode."
- https://www.win.tue.nl/~aeb/linux/uc/nfc_vs_nfd.html. "Unicode Normalization."
- https://en.wikipedia.org/wiki/Unicode_equivalence. "Unicode Equivalence."

As always, we remain open to developing AGNT and ANLEX in ways that are most useful to the needs of students and readers of God's Word.

Thank you for your continued support of *The AGNT Project*, for faithfully marketing the AGNT and ANLEX databases, and for making these state-of-the-art tools for studying the Greek New Testament available to students, scholars, pastors, translators, and laymen worldwide.

John Hughes
Agent for *The AGNT Project*
johnhughes@centurytel.net
Phone: 406.862.7289
FAX: 406.862.0917