SyrCOM-99: Proceedings of the Third International Forum on Syriac Computing, pp. 12-18

Encoding Syriac in ISO/IEC 10646 (Unicode)

SARGON HASSO
shasso@lucent.com
Lucent Technologies

GEORGE KIRAZ
gkiraz@research.bell-labs.com
Bell Labs, Lucent Technologies

PAUL NELSON
paulnel@microsoft.com
Microsoft Corporation

Abstract

We present a description of a proposal submitted to Unicode, Inc. to encode Syriac scripts in ISO/IEC 10646. The goal is to produce a uniform Syriac character-encoding scheme that provides a standard method for writing current Syriac and a common means for the electronic storage and interchange of manuscript data. The languages and dialects that employ the Syriac scripts in one form or another, and either currently or sometimes during the past include, among others: Literary Syriac, Neo-Aramaic, Garshuni, and Christian-Palestinian Aramaic. This proposal encodes most characters used in Syriac texts that encompass three major scripts: Estrangela, Serto or West Syriac, and East Syriac scripts. This proposal can also serve as a specification document for implementers interested in developing a multi-lingual software systems, e.g., word processing, desktop publishers, database systems, dictionaries, and others.

I. INTRODUCTION

This paper describes a proposal that we submitted to Unicode Consortium to include the Syriac scripts in the Unicode Standards. This proposed assignment for Syriac characters was created by merging a previous proposal worked on by Sargon Hasso, Rick McGowan, and Michael Everson with a proposal created by George Anton Kiraz and Paul Nelson. The proposal of characters to use has been derived from characters currently used in writing the Syriac language and characters commonly found in a broad range of manuscripts used for study. The desire is to be able to not only provide a standard method for writing current Syriac, but also a common means for the electronic storage and communication of manuscripts and various textual data. The previous proposals deemed incomplete for approval as a standard. Therefore, the current proposal augmented previous attempts in the following areas:

Vowel, overstrike, and punctuation marks were added to allow for the proper writing of languages which employ the Syriac type styles.

Additional characters were added to allow for the writing of non-Syriac languages that employ Syriac type styles, e.g. Christian Palestinian Aramaic (also known as Palestinian Syriac), Garshuni (Arabic written in Syriac type styles) and various modern Neo-Aramaic dialects.

Joining rules. Letters were classified into joining classes (e.g., which letters connect to their neighboring letters and in which manner). Special joining rules for the Alaph were indicated.

Ligature rules were added to show how Syriac should handle the combining of letters into ligatures.

Abbreviation rules were stated to show how the Syriac abbreviation is implemented.

In section II we discuss very briefly the Unicode Standard and its relation to the International Standard ISO/IEC 10646. To better appreciate what the proposal is about, we give in section III a brief overview of the Syriac Language, its various dialects, and the Syriac scripts it uses. In section IV, we examine in some details the contents of the proposal. The proposal is published on the World Wide Web (Nelson, Kiraz, Hasso: 1998).

II. The Unicode Standard and ISO/IEC 10646

The Unicode Consortium is a standards organization body that was incorporated in 1991 and run under the name Unicode, Inc. Its purpose is to promote a unified standard, to aid in its implementation, and to maintain control over its future maintenance. The Unicode Standard is a universal character encoding scheme for character representation in computer processing and information exchange. It is based on the widely common ASCII encoding scheme but has the potential to encode almost all the written scripts in the word. It accomplishes this by assigning a unique 16-bit wide number to each character in every script.

Each and every character in the Unicode Standard has a unique identification number referred to as code value, and a unique character name that serves as its definition, referred to as name. Code values are specified in hexadecimal number format following the designation "U". For example, the SYRIAC LETTER Kaph, has code value U+071F (1823 in decimal format). The Unicode Standard does not concern itself of what the representational form these characters have. Technically, the character code serves as an abstract unit, and its representation form is called a glyph. Information processing applications deal internally with codes. When they display information, they map these codes to their glyphs. The Unicode standard also provides some additional information that help application developers implement these encoding schemes especially those applications that handle complex multilingual scripts. Thus application software will be able to process multilingual textual information with ease by referring to character encoding databases.

The Unicode Consortium encourages the submissions of new scripts for possible inclusion in the Unicode Standard. Our work on Syriac script over the past few years has culminated in a fairly complete standard document nearing its full approval as part of the growing Unicode Standard. The Unicode Standard is fully compatible, code-for-code, with International Standard ISO/IEC 10646 (Unicode 1996).

III. The Syriac Language and Syriac Scripts

The Syriac language belongs to the Aramaic family of languages. The earliest datable Syriac writing is in the form of inscriptions from Birecik, dating A.D. 6 (Maricq 1962, Pirenne 1963). Three legal documents from the third century (dated 28 Dec 240, 1 Sept 242 and 243, respectively) were discovered in the Euphrates valley (Brock 1991, Drijvers 1972). The earliest literary Syriac manuscript is dated November A.D. 411 (Hatch, 1946) with an unbroken tradition of writing till the modern time.

Today, Syriac is the active liturgical language of many communities in the Middle East (Syrian Orthodox, Assyrian, Maronite, Syrian Catholic and Chaldaean) and Southeast India (Syro-Malabar and Syro-Malankara). It is also (in various dialectical forms written using the Syriac scripts) the native language of a million or two (although no reliable statistics can be found). Syriac is widely used among its natives in their native lands, as well as the diaspora in Europe, the Americas and Australia. Additionally, Syriac is the subject of study for many Western scholars who publish texts in the Syriac scripts, e.g., the monumental ca. 230-volume Scriptores Syri of the CSCO series.

Syriac is divided into two dialects. West Syriac is used by the Syrian Orthodox, Maronites and Syrian Catholics. East Syriac is used by the Assyrians (i.e., Ancient Church of the East) and Chaldaeans. The two dialects are very similar in grammar and vocabulary. They differ, however, in phonology (i.e., pronunciation) which has no impact on this work. However, each of the two dialects has its own script, each script with its own idiosyncrasies.

There are a number of other languages and dialects that employ the Syriac script in the modern time in one form or another. These are:

Literary Syriac. The primary usage of Syriac scripts.

Neo-Aramaic dialects. "Modern Aramaic languages," Hobermann notes "have been written with the Syriac, Hebrew, Cyrillic, and Roman scripts, but only the Syriac script has gained widespread use." (Hobermann, 1996, p. 504). To this category of languages belong a number of Eastern Modern Aramaic dialects known as Swadaya (also called 'vernacular Syriac', 'modern Syriac', 'modern Assyrian' etc., spoken mostly by the Assyrians and Chaldaeans of Iraq, Turkey and Iran), and the Central Aramaic dialect of Turoyo (spoken mostly by the Syrian Orthodox of the Tur Abdin region in Southeast Turkey). These formerly "spoken" dialects have become literary in the past hundred years or so (see Murre-van den Berg,

1994). They employ the Syriac scripts in addition to overstrike marks to indicate sounds not found in, but similar to, Syriac ones (Maclean 1971).

Garshuni, i.e., Arabic written in the Syriac script. This mode of writing is currently used for writing Arabic liturgical texts amongst the Syriac-speaking Christians. A large corpus of manuscripts ranging from the 8th century till the modern day exists in Garshuni (Mingana 1933). Garshuni employs two additional letters and the Arabic set of vowels and overstrike marks.

Christian Palestinian Aramaic (known also as Palestinian Syriac) employs the Syriac scripts with one additional letter, the reversed Pe (Schulthess, 1979). This dialect is no longer spoken, but there has been recent scholarly interest in publishing its texts, e.g. (Müller-Kessler and Sokoloff, 1996 and 1997).

Other languages. The Syriac scripts were used at various historical periods for writing Armenian and some Persian dialects. Syriac-speakers employed them for writing Arabic, Ottoman Turkish, and Malayalam. Manuscripts written in this manner survive and are the subject of study by Western scholars. Syriac-speaking peoples wrote Ottoman Turkish in the Syriac scripts as late as the beginning of this century (e.g., Al-Intibah newspaper published in New York by immigrants in the 1900s-1920s). They continue to write Arabic in this manner (see under Garshuni above).

Syriac texts employ the following scripts:

Estrangela script. Estrangela (a word derived from Greek strongulos meaning 'rounded') is the oldest script. Ancient manuscripts use this writing style exclusively. Estrangela has seen a revival in the twentieth century (it has seen an earlier revival in the 10th century whence it has been defunct for a hundred years, as we are told by the historian Bar Ebroyo, (Hatch 1946, p. 26). Estrangela is used today in West and East Syriac texts for writing headers, titles and subtitles. It is also used in cards, engravings, etc. Most importantly, this script is the current standard in writing Syriac texts amongst Western scholarship, almost exclusively.

Serto or West Syriac script (also misnamed "Jacobite"). This script is the most cursive of all. It emerged around the 8th century (Healey 1990) and is used today in West Syriac texts, as well as Turoyo (Central Neo-Aramaic) and Garshuni.

East Syriac script (also misnamed "Nestorian"). Its early features appear as early as the sixth century; it developed into its own script by the 12th or 13th centuries (Healey 1990). It is used today for writing East Syriac texts, as well as Swadaya (Eastern Neo-Aramaic). It is also used today in West Syriac texts for headers, titles and subtitles alongside the Estrangela script.

Christian Palestinian Aramaic. Manuscripts of this dialect employ a script that is akin to Estrangela. Indeed, it can be considered a sub-category of Estrangela.

This proposal provides for usage of the scripts mentioned above. Additionally, it provides for letters and diacritics used in Neo-Aramaic languages, Christian Palestinian Aramaic, and Garshuni languages.

IV. A Closer Look at the Syriac Proposal Contents

In this section we examine briefly some of the contents of this proposal. This can serve as an introduction to the understanding of the complete proposal. The Unicode Standard has three basic design principles in which any character set must have (Unicode 1999):

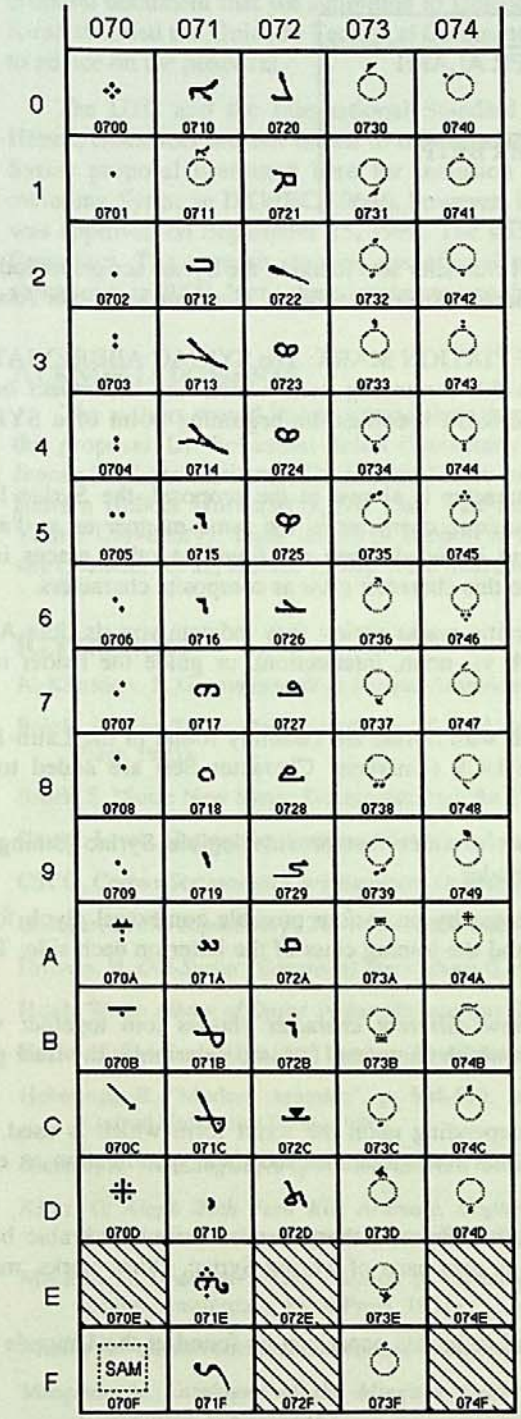
define the smallest useful elements of text to be encoded,

assign a unique code to each elements, and

provide basic rules for encoding and interpreting text so that programs can successfully read and process text.

Our proposal satisfies all of these design requirements. Additionally, we felt that we needed to address issues about the Syriac scripts not explicitly examined elsewhere. With this in mind, we hope we can serve the Syriac-speaking communities, publishers, scholars, and information processing experts with all the information they need to use, store, and exchange Syriac textual information in electronic or paper-published form. The Unicode Standard groups related characters in code blocks. Each code block is

represented by its code chart. The complete code chart for Syriac script looks like this (characters are in Estrangela script):



The code for each character is computed by appending the column heading numerical value with the character's row designation. For example, the code of Alaph is 0710 (071 is the column value and 0 is the row value) and the code of Taw is 072D. For ease of usage, the codes are specified in each entry in the above chart.

The Unicode standard also requires, as was mentioned previously, character names that serve as their definitions. Here is a sample of the Syriac name chart:

Glyph	Unicode	Name
ܐ	U+0710	SYRIAC LETTER ALAPH
ܒ	U+0712	SYRIAC LETTER BETH

The code chart and its companion name chart cover all character sets found in the Syriac scripts including:

1. *Letters of the Alphabet.* These are the base characters used for Syriac, Garshuni and other Aramaic dialects.

Syriac Control Character - The SYRIAC ABBREVIATION MARK: The SYRIAC ABBREVIATION MARK (SAM) U+070F is a user-selectable zero-width formatting code which has no impact on the shaping process of Syriac characters. The use of the SAM specifies the beginning point of a SYRIAC ABBREVIATION.

Combining Characters. Only one combining character is shown in the proposal--the Syriac Letter Yudh-He, U+071E. This combination is used as a unique character in the same manner as an "æ". A number of combining diacritics unique to Syriac are included when not found at other places in the Unicode Standard. The Unicode Standard also refers to this character class as composite characters.

Diacritic Marks/Vowels. The function of the diacritic marks varies: they indicate vowels (like Arabic and Hebrew), mark grammatical attributes (e.g., verb vs. noun, interjection), or guide the reader in the pronunciation/reading of the given text.

Punctuation Marks. Most punctuation marks used with Syriac are currently found in the Latin-1 and Arabic blocks. Punctuation marks not found in the UCS (Universal Character Set) are added to this proposal.

The proposal document provides a minimum set of rules that provide legible Syriac joining and ligature substitution behavior. These specifications include:

Joining Classes. Each Syriac character is represented by up to four possible contextual glyph forms. The form used is determined by the its joining class and the joining class of the letter on each side. These classes are identical to those outlined for Arabic.

Joining Rules. These rules specify explicitly how different character classes join together when rendered (displayed) visually. Naturally, the context in which characters fall into determines the final glyph form.

Ligature Classes. Ligatures are valid in Syriac depending upon the script form which is used. The proposal lists most common character combinations that form ligatures and their specification as either optional or mandatory for rendering purposes.

The proposal contains a chart that specifies which characters that are used from the Arabic block: punctuation marks, tatweel and the shaddah are used as core parts of writing Syriac. Other marks, mainly diacritics, are used in Garshuni.

The complete proposal with all the charts referred to in this paper can be found at the Unicode Inc., World Wide Web site (Nelson, Kiraz, Hasso: 1998).

V. Conclusion

The Unicode Standard is worldwide character encoding scheme that covers just about all the written world scripts. Unicode allows for all textual elements to be encoded in a universal, efficient, uniform, and unambiguous manner. It uses 16-bit numbers that has the capacity to encode over 65,000 characters. The Unicode Standard, Version 2.0 contains 38,885 character codes. Many world scripts are being submitted

and reviewed for possible inclusion in the standard.

Our work, with assistance of many of our colleagues over the past few years, has culminated in a proposal document that we submitted to Unicode in 1998. On February 27, 1998, Paul Nelson and George Kiraz attended the Unicode Technical Committee (UTC) meeting, held at Microsfot's Offices in California, to advice on the proposal.

The UTC and the International Standard Organization (ISO) collaborate in assigning code points. Hence, characters are not added to one standard without being approved for the other. UTC approved the Syriac proposal presented here for inclusion in Unicode on February 27, 1998. The ISO process for including Syriac in ISO/IEC 10646, however, is much slower and consists of seven stages. The fifth stage was approved on September 25, 1998. The sixth stage is holding a final ballot (and at this stage is a mere formality). The seventh stage is the official publication of the proposal. This is expected to take place sometimes in 2000. Until then, implementing the Syriac block of Unicode is at implementers risk!

Acknowledgments

The authors would like to extend their thanks to a number of people who reviewed and commented on this proposal: Dr. Sebastian Brock (University of Oxford), Dr. J. Coakley (Harvard University), Dr. K.D. Jenner (Peshitta Institute), Dr. Heleen Murre-van den Berg (Leiden University), Dr. Edward Odisho (North Eastern Illinois University), Dr. Luk Van Rompay (Leiden University) and Dr. Jan Wilson (Brigham Young University). Their letters of support to the Unicode Consortium, which are available from the Web site (Nelson, Kiraz, Hasso: 1998), were instrumental in getting the proposal through.

References

Al-Kfarnissy, P. *Grammaire de la Langue Araméenne Syriaque* (Beyrouth, 2nd edition, 1962).

Brock, S. "The Syriac Background", p. 30, n. 2. In Lapidge, M. (ed.), *Archbishop Theodore* (Cambridge University Press, 1995).

Brock, S. "Some New Syriac Documents from the Third Century AD", Aram 3 (1991), pp. 259-67.

Costaz, Lewis, *Grammaire Syriaque* (Beyrouth: Librairie Orientale, 1955).

CSCO. *Corpus Scriptorum Christianorum Orientalium. Scriptores Syri* (Lovanii: 1904-1998).

Diringer, D. *The Alphabet*, p. 218 (New York: Funk & Wagnalls, 1968).

Drijvers, H. *Old-Syriac (Edessean) Inscriptions* (Leiden: Brill, 1972).

Hatch, W. *An Album of Dated Syriac Manuscripts* (Boston: The American Academy of Arts and Sciences, 1946).

Healey, J. *The Early Alphabet* (London: The British Museum, 1990).

Hoberman, R. "Modern Aramaic", p. 504-510. In Daniels, P. and Bright, W. (eds) *The World's Writing Systems* (Oxford University Press, 1996).

Katzner, K. *The Languages of the World* (New York: Funk & Wagnalls, 1975).

Kiraz, G. *Alaph Beth Font Kit: Aramaic, Coptic, Hieroglyphic, North Semitic, Phoenician, Sabaean, Syriac and Ugaritic Fonts*. (Los Angeles: Alaph Beth Computer Systems).

Maclean, A. *Grammar of the Dialects of Vernacular Syriac as Spoken by the Eastern Syrians of Kurdistan...*, 2nd edition (Amsterdam: Philo Press, 1971).

Maricq A. "La plus ancienne inscription syriaque: cell de Birecik" *Syria* 39 (1962), pages 88-100.

Mingana, A. *Catalogue of the Mingana Collection of Manuscripts*, vol. 1, Syriac and Garshuni Manuscripts (Cambridge, U.K., 1933).

Müller-Kessler, C. and Sokoloff, M. *A Corpus of Christian Palestinian Aramaic* (Groningen: Styx, 1996-1997).

Murre-van den Berg, H. *From a Spoken to a Written Language, The Introduction and Development of Literary Urmia Aramaic in the Nineteenth Century*, Ph.D. thesis, Leiden University, 1994.

Nelson, P., Kiraz, G., Hasso, S.. *Proposal to encode Syriac in ISO/IEC 10646*, 1998, (URL:

http://www.unicode.org/pending/syriac/default.htm, 1998).

Noldeke, T. *Compendious Syriac Grammar*, translated by James Crichton (London: 1904).

Oudo [sic.], T. *Treasure of the Syriac Language* (Holland: Bar Hebraeus Verlag, 1985).

Pirenne, J. "Aux origines de la graphie syriaque" in *Syria* 40 (1963).

Robinson, T. *Paradigms and Exercises in Syriac Grammar* (Oxford University Press, 4th edition, 1978).

Schulthess, Friedrich, *Lexicon Syropalaestinum* (Amsterdam: APA Oriental Press, 2nd edition, 1979). [Material on Pe, pp.154 ff.]

Segal, J., *The Diacritical Point and the Accent in Syriac* (Oxford University Press, 1953).

Smith, Payne, *Thesaurus Syriacus* (Oxford: The Clarendon Press, 1879).

Unicode, Inc., *Unicode Standard, Version 2.0* (Addison Wesley Longman, 1996).

Unicode, Inc., *The Unicode Standard, A Technical Introduction* (URL: http://www.unicode.org/unicode/standard/principles.html, 1999).