New Canonical decomposition and composition processes for Hangeul

Kyongsok KIM, Professor,
Division of Computer Science and Engineering,
Busan National University, Busan 609-735, South KOREA

Internet: gimgs@HANGEUL.pnu.edu
Phone: Home/ +82-51-947-6581, Office/ +82-51-510-2292
Fax: +82-51-515-2208

Abstract

In this paper, the author proposes new canonical decomposition process (NFD: Normalization Form D) and composition processes (NFC: Normalization Form C) for Hangeul, Korean script. The new processes can handle both Modern and Old Hangeul in a consistent way. Therefore, the new processes are more general than the ones described in UAX #15, which handle only Modern Hangeul syllables.

Since NFD and NFC will play an important role in sorting and searching, Internet-related applications such as Multilingual Domain names, databases, etc., we need to establish good NFD and NFC for Hangeul before they are widely used.

Keywords:

1) Hangeul, Hangul, Korean
2) canonical, decomposition, composition
3) normalization, form
4) Multilingual domain name, Internet, sort, search

1. Introduction

In Section 1, we will see a brief introduction of Hangeul, Korean script, and terms used in this paper. In Section 2, we will review previous works about Hangeul Normalization Forms, especially four Normalization Forms described in UAX #15 [UAX15].

Then, in Section 3, we will propose new canonical decomposition process (NFD: Normalization Form D) and composition processes (NFC: Normalization Form C) for Hangeul, which are better than the previous proposals.

Conclusions are described in Section 4.

Now let's see a brief introduction of Hangeul, Korean script, and terms used in this paper.

Note 1. The spelling of "Hangeul" follows ISO TS (Technical Specification) 11941, Information and Documentation -- Transliteration of Korean script into Latin characters. First edition, 1996-12-01.

Note 2. The alphabetical order of Hangeul letters follows the one adopted in South Korea.

1) Hangeul

It is the name of Korean alphabet or Korean script. 'Han' means great and 'geul' means a script.

2) Modern Hangeul simple letters

Modern Hangeul syllable-initial, syllable-peak, and syllable-final simple letters are shown below (U+xxxx indicates UCS-2 code positions [ISO10646]).

a) Syllable-initial simple letters (14):

ᄀ U+1100, ᄂ U+1102, ᄃ U+1103, ᄅ U+1105, ᄆ U+1106, ᄇ U+1107, ᄉ U+1109, ᄋ U+110B, ᄌ U+110C, ᄎ U+110E, ᄏ U+110F, ᄐ U+1110, ᄑ U+1111, ᄒ U+1112

b) Syllable-peak simple letters (10):

ᅡ U+1161, ᅣ U+1163, ᅥ U+1165, ᅧ U+1167, ᅩ U+1169, ᅭ U+116D, ᅮ U+116E, ᅲ U+1172, ᅳ U+1173, ᅵ U+1175

c) Syllable-final simple letters (14):

ᆨ U+11A8, ᆫ U+11AB, ᆮ U+11AE, ᆯ U+11AF, ᆷ U+11B7, ᆸ U+11B8, ᆺ U+11BA, ᆼ U+11BC, ᆽ U+11BD, ᆾ U+11BE, ᆿ U+11BF, ᇀ U+11C0, ᇁ U+11C1, ᇂ U+11C2

3) Modern Hangeul complex letters

Two or three simple letters can be combined to form a complex letter.

Modern Hangeul syllable-initial, syllable-peak, and syllable-final complex letters are shown below.

a) Syllable-initial complex letters (5):

ᄁ U+1101, ᄄ U+1104, ᄈ U+1108, ᄊ U+110A, ᄍ U+110D

b) Syllable-peak complex letters (11):

ᅢ U+1162, ᅤ U+1164, ᅦ U+1166, ᅨ U+1168, ᅪ U+116A, ᅫ U+116B, ᅬ U+116C, ᅯ U+116F, ᅰ U+1170, ᅱ U+1171, ᅴ U+1174

c) Syllable-final complex letters (13):

ᆩ U+11A9, ᆪ U+11AA, ᆬ U+11AC, ᆭ U+11AD, ᆰ U+11B0, ᆱ U+11B1, ᆲ U+11B2, ᆳ U+11B3, ᆴ U+11B4, ᆵ U+11B5, ᆶ U+11B7, ᆹ U+11B9, ᆻ U+11BB

4) Modern syllable-initial, syllable-peak, and syllable-final letters

Modern Hangeul syllable-initial, syllable-peak, and syllable-final letters, including both simple and complex letters, are shown below in alphabetical order of South Korea:

a) Syllable-initial letters (19):
ᄀ U+1100, ᄁ U+1101, ᄂ U+1102, ᄃ U+1103, ᄄ U+1104, ᄅ U+1105, ᄆ U+1106, ᄇ U+1107, ᄈ U+1108, ᄉ U+1109, ᄊ U+110A, ᄋ U+110B, ᄌ U+110C, ᄍ U+110D, ᄎ U+110E, ᄏ U+110F, ᄐ U+1110, ᄑ U+1111, ᄒ U+1112

b) Syllable-peak letters (21):
ᅡ U+1161, ᅢ U+1162, ᅣ U+1163, ᅤ U+1164, ᅥ U+1165, ᅦ U+1166, ᅧ U+1167, ᅨ U+1168, ᅩ U+1169, ᅪ U+116A, ᅫ U+116B, ᅬ U+116C, ᅭ U+116D, ᅮ U+116E, ᅯ U+116F, ᅰ U+1170, ᅱ U+1171, ᅲ U+1172, ᅳ U+1173, ᅴ U+1174, ᅵ U+1175

c) Syllable-final letters (27):
ᆨ U+11A8, ᆩ U+11A9, ᆪ U+11AA, ᆫ U+11AB, ᆬ U+11AC, ᆭ U+11AD, ᆮ U+11AE, ᆯ U+11AF, ᆰ U+11B0, ᆱ U+11B1, ᆲ U+11B2, ᆳ U+11B3, ᆴ U+11B4, ᆵ U+11B5, ᆶ U+11B7, ᆷ U+11B7, ᆸ U+11B8, ᆹ U+11B9, ᆺ U+11BA, ᆻ U+11BB, ᆼ U+11BC, ᆽ U+11BD, ᆾ U+11BE, ᆿ U+11BF, ᇀ U+11C0, ᇁ U+11C1, ᇂ U+11C2

5) A complete Hangeul syllable: 2-letter and 3-letter syllable

In a written text, a word is a sequence of Hangeul syllables, most of which are complete syllables. Each syllable is written in a square.

There are two types of complete' Hangeul syllables:

5-a) A 2-letter syllable is composed of syllable-initial and syllable-peak (i.e., vowel) letters, where each letter can be either simple or complex.

5-b) A 3-letter syllable is composed of syllable-initial, syllable-peak (i.e., vowel), and syllable-final letters, where each letter can be either simple or complex.

6) An incomplete Hangeul syllable

An 'incomplete' syllable is a syllable which is not complete. There are four types of incomplete syllables: i) a syllable-initial letter alone, ii) a syllable-peak letter alone, iii) a syllable-final letter alone, iv) syllable-peak and -final letters alone.

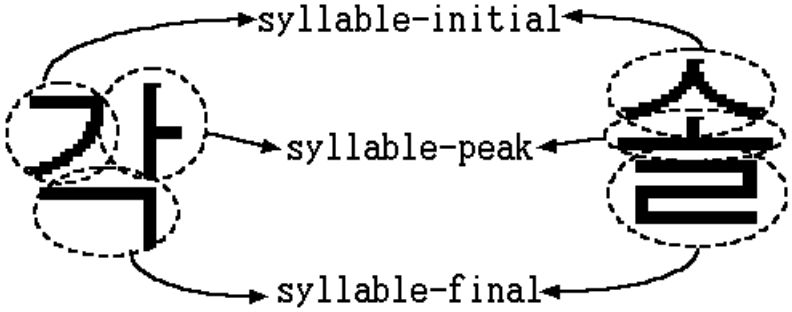
Examples: i) ᄀ (U+1100 1160), ii) ᅡ (U+115F 1161), iii) ᆨ (U+115F 11A8), iv) 니 (U+115F 1175 11AB)

7) Consonantal letters

In Hangeul, consonantal letters are classified into syllable-initial and -final letters depending on the physical position in a square for a syllable.

A syllable-initial letter is a consonantal letter appearing to the left and/or above a syllable-peak letter, whereas a syllable-final letter is a consonantal

letter appearing below syllable-initial and -peak letters.



8) Old Hangeul simple letters

Old Hangeul simple letters used for Old Hangeul are shown below:

a) Old syllable-initial simple letters (3): ㅿ U+1140, ㆁ U+114C, ㆆ U+1159

b) Old syllable-peak simple letter (1): ㆍ U+119E

c) Old syllable-final simple letters (3): ㅿ U+11EB, ㆁ U+11F0, ㆆ U+11F9

In addition, there are six Old syllable-initial simple letters used for foreign words.

8-d) U+113C, U+113E, U+114E, U+1150, U+1154, U+1155

9) Old Hangeul complex letters

There are many Old complex letters. 158 Old complex letters are included in ISO/IEC 10646. There seem fairly many (tens of ?) Old complex letters not included in ISO/IEC 10646.

10) Hangeul letters in ISO/IEC 10646-1

238 Hangeul letters (whose code positions are U+11xx) included in ISO/IEC 10646-1 are classified depending on whether they are Modern or Old, whether they are simple or complex, and whether they are syllable-intial, -peak, or -final. A summary table of letters is shown below:

Table 1. Classification of Hangeul letters included in ISO/IEC 10646-1.

	Modern			Old			Modern + Old		
	simple	complex	sub-tot.	simple	complex	sub-tot.	simple	complex	tot.
syllable-initial	14	5	19	9	62	71	23	67	90
syllable-peak	10	11	21	1	44	45	11	55	66
syllable-final	14	13	27	3	52	55	17	65	82
total	38	29	67	13	158	171	51	187	238

11) Hangeul Filler characters

Filler characters are used for representing incomplete syllables. There are two Filler characters.

syllable-initial Filler character: U+115F
syllable-peak Filler character : U+1160

Usage of Filler characters are shown below:
- ᄀ : U+1100 1160 (syllable-initial letter alone: append U+1160)
- ᅡ: U+115F 1161 (syllable-peak letter alone: prepend U+115F)
- ᆨ: U+115F 11A8 (syllable-final letter alone: prepend U+115F)

12) Old Hangeul tone marks: Bangjeom

Old Hangeul tone marks are used for representing the tones of syllables. There are two tone marks.

single dot tone mark: ◌ U+302E
double dot tone mark: ◌ U+302F

13) 2-complex and 3-complex letters

A 2-complex letter is a complex letter composed of two simple letters. Likewise, a 3-complex letter is a complex letter composed of three simple letters.

Examples of complex letters are shown below:
2-complex letters: ᄁ U+1101, ᅪ U+1162, ᆫᄒ U+11AD
3-complex letters: ᅶᅫ U+116B, 궤 U+1170

14) IPF (Initial-Peak-Final) and Wanseong forms for Hangeul syllables

Modern Hangeul syllables can be represented either in IPF (initial-peak-final) form or in Wanseong (precomposed) form.

For example, syllable "가" can be represented as follows:

a) in IPF form: U+1100 1161

b) in Wanseong form: U+AC00

In contrast, Old Hangeul syllables must be represented in IPF form, since no Old Hangeul syllables in Wanseong form are included in ISO/IEC 10646-1.

For example, "Old Hangeul syllable "ᄀᆡ" can be represented as follows:

a) in IPF form: U+1100 119E

b) in Wanseong form: the syllable cannot be represented in Wanseong form.

Table 2. Possibilities of representing Modern and Old Hangeul syllables in IPF or Wanseong form.

	in IPF form	in Wanseong form
Modern Hangeul	syllables composed using 67 Modern letters	11,172 syllables
Old Hangeul	syllables composed using 238 Modern and Old letters	(NO Old syllables included in ISO/IEC 10646-1)

15) 94 compatibility letters in ISO/IEC 10646-1 [U+3131 - 318E]

The 94 letters in the range of U+3131 - 318E are included for compatibility. They are from KS X 1001 (formerly, KS C 5601). Compatibility means that these letters can be used when converting Old files in KS X 1001 to files in ISO/IEC 10646. Completely new files are not supposed to be represented in compatibility letters; instead, they should be represented in U+11XX or U+AC00-D7A3.

Now let's see the usage of these letters.

a) Each of these 93 letters (excluding the FILLER character: U+3141) can represent an independent letter.

Examples: ㄱ (U+3131), ㄱㄴㄷ (U+3131 3134 3137), ㅏ (U+314F), ㄶ (U+3136).

b) Usage of FILLER character (U+3164): it has two different functions.
A sequence of four code positions can represent one syllable as follows:
FILLER (U+3164),
consonantal letter (U+3131-314E, U3165-3186),
vowel letter (U+314F-3163, U+3187-318E), and
consonantal (U+3131-314E, U3165-3186) letter or FILLER (U+3164).

FILLER indicates that the FILLER character and following three code positions together represent one syllable, not three independent letters.

FILLER (U+3164) before U+3131 in example c), (1) and the ones before U+3138 or 3131 in example c), (2) fall in this category.

In case of 2-letter syllable, FILLER is included at end to indicate that there is no syllable-final letter. FILLER (U+3164) after U+314F in example c), (1) below falls in this category.

c) A syllable, especially a syllable which is not one of the 2,350 syllables in KS X 1001, can be represented as follows:

(1) A 2-letter syllable can be represented in the form of FILLER (U+3164), consonantal letter, vowel letter, and FILLER (U+3164).

Example: 가 (U+3164 3131 314F 3164)

(2) A 3-letter syllable can be represented in the form of FILLER (U+3164), consonantal letter, vowel letter, and consonantal letter.

Examples: 뚬 (U+3164 3138 3157 3141). This syllable is not one of 2,350 syllables in KS X 1001.

강 (U+3164 3131 314F 3147)

d) "ㄱ ㅏ" (two independent letters) and "가" (one syllable) are represented differently as follows:

- ㄱ ㅏ: U+3131 314F (Two letters, not one syllable)
- 가: U+3164 3131 314F 3164 (One syllable. The first FILLER (U+3164 before U+3131) indicates that the FILLER and following three code positions together represent one syllable, not three letters. The last Filler (U+3164 after U+314F) indicates that this syllable is a two-letter syllable (i.e., no syllable-final letter).

2. Previous Works

In UAX #15 [UAX15], four normalization forms are described:

1) Normalization Form D (NFD): canonical decomposition
2) Normalization Form C (NFC): canonical decomposition, followed by canonical

composition
3) Normalization Form KD (NFKD): compatibility decomposition
4) Normalization Form KC (NFKC): compatibility decomposition, followed by canonical composition

Once a canonical decomposition process for Hangeul is defined, compatibility decomposition process for Hangeul can be easily defined. Therefore, in this paper, we will focus on canonical decomposition and composition processes for Hangeul, and as a result, Normalization Forms D (NFD) and C (NFC) for Hangeul.

2.1 Canonical Decomposition and Canonical Composition Processes

We may normalize Unicode-encoded text to one particular sequence. We can normalize into one of two forms [Unic30]:

1) systems that cannot handle nonspacing marks can normalize to precomposed characters; ....
2) In systems that can handle nonspacing marks, it may be useful to normalize so as to eliminate precomposed characters. ...

Process 1) above is NFD and process 2) above is NFC.

A character that is equivalent to a sequence of one or more other characters is called a decomposable character, a precomposed character, or a composite character.
A decomposition of a decomposable character is a sequence of one or more characters that is equivalent to a decomposable character.

2.2 Hangeul syllable decomposition process in UAX #15

The canonical decomposition process is described in [UAX15]. For a given Modern Hangeul syllable, a sequence composed of two or three of the 67 Modern Hangeul letters shown in 4), Section 1, is derived, which is a canonical decomposition. The sequence is composed of Modern Hangeul syllable-initial, syllable-peaK, and optionally syllable-final letters. Whenever possible, a complex letter is used in place of two or three simple letters.

Example in [UAX15]:
- An original syllable: 갃 (U+AC03)
  --> [NFD]: ᄀ + ᅡ + ᆪ (U+1100 1161 11AA) --> [NFC]: 갃 (U+AC03)

Note. A 2-complex letter ᆪ (U+11AA) is not decomposed any further, although it could be decomposed into two simple letters ᆨ (U+11A8) and ᆺ (U+11BA).

To help readers understand, we made up more examples:

강 (U+AC15) --> [NFD] ᄀ + ᅡ + ᆼ (U+1100 1161 11BC)
갉 (U+AC0C) --> [NFD] ᄀ + ᅡ + ᆱ (U+1100 1161 11B0)
Note. ᆱ (U+11B0) is not decomposed any further.
객 (U+AC1D) --> [NFD] ᄀ + ᅫ + ᆨ (U+1100 1162 11A8)
Note. ᅫ (U+1162) is not decomposed any further.
가 (U+AC00) --> [NFD] ᄀ + ᅡ (U+1100 1161)

2.3 Hangeul syllable composition in UAX #15

The composition process is described in [UAX15]. For a given canonical decomposition of a Modern Hangeul syllable, we can find a syllable whose code position is in the range of U+AC00 - D7A3. That syllable is a composition in [UAX15].

Note. In the Unicode 3.0 [Unic30], this composition is not described explicitly as "canonical" composition.

Example in [UAX15]:
- An original syllable: 갃 (U+AC03)
--> [NFD] ᄀ + ᅡ + ᆪ (U+1100 1161 11AA) --> [NFC] 갃 (U+AC03)

To help readers understand, we made up more examples:
ᄀ + ᅡ + ᆼ (U+1100 1161 11BC) --> [NFC] 강 (U+AC15)
ᄀ + ᅡ + ᆱ (U+1100 1161 11B0) --> [NFC] 갉 (U+AC0C)
ᄀ + ᅫ + ᆨ (U+1100 1162 11A8) --> [NFC] 객 (U+AC1D)
ᄀ + ᅡ (U+1100 1161) --> [NFC] 가 (U+AC00)

3. New Canonical decomposition and composition processes for Hangeul

Although Normalization Forms KD (NFKD) and KC (NFKC) for Hangeul are not treated in this paper, based on new NFD and NFC, it should be straightforward to define new NFKD and NFKC. Therefore, in this paper, we will focus on new canonical decomposition and composition processes for Hangeul, and as a result, new Normalization Forms D (NFD) and C (NFC) for Hangeul.

3.1 The main idea of new Hangeul decomposition/composition processes

Canonical decomposition and composition processes for Hangeul, as shown in UAX #15, have limitations.

1) UAX #15 decomposition process only decomposes 11,172 Modern Hangeul. Therefore, UAX #15 decomposition process cannot decompose complex (both Modern and Old) letters.

2) UAX #15 composition process only composes Modern letters into syllables. Therefore, UAX #15 composition process cannot compose letters into Old complex letters.

Therefore, we propose new canonical decomposition and composition processes for Hangeul, Korean script. These new processes solve the limitations of UAX #15 decomposition/composition.

A new decomposition process decomposes Hangeul letters/syllables into simple letters. Then a new composition process composes simple letters into complex letters (whenever possible), but not into syllables. These are the main idea of new decomposition and composition processes.

Tables 3 and 4 compare UAX #15 decomposition/composition and new ones.

Table 3. Comparison of UAX #15 Decomposition and a new Decomposition.

Decomposition	input	output
UAX #15	11,172 Modern syllables only	1) Modern complex letters (whenever possible) 2) Modern simple letters
A New process	1) Modern syllables 2) Modern complex letters 3) Old complex letters	1) Modern simple letters 2) Old simple letters

Table 4. Comparison of UAX #15 Composition and a new Composition.

Composition	input	output
UAX #15	1) Modern complex letters 2) Modern simple letters	11,172 Modern syllables
A New process	1) Modern simple letters 2) Old simple letters	1) Modern complex letters (whenever possible) 2) Modern simple letters 3) Old complex letters (whenever possible) 4) Old simple letters

3.2 A new decomposition process for Hangeul

A new decomposition process decomposes syllables and complex (both Modern and Old) letters as follows. To decompose a syllable, start at step D1; to decompose a complex letter, start at step D2.

step D1) For each Hangeul syllable in Wanseong form, decompose a syllable into syllable-initial, syllable-peak, and possibly syllable-final letters as described in [ISO10646].

step D2) For each complex letter, decompose a complex letter into simple-letter-1, simple-letter-2, and possibly simple-letter-3. For information about decomposing a complex letter into simple letters, see Appendix 1.

step D3) The result is the canonical decomposition of the given Hangeul syllable or complex letter. Note that the canonical decomposition has only simple letters and possibly Filler characters and tone marks, but neither syllables in Wanseong form nor complex letters.

First, we will review the example in [UAX15]:
- An original syllable: 갃 (U+AC03)
  --> [NFD] ᄀ + ᅡ + ᆪ (U+1100 1161 11AA)

As noted earlier, a 2-complex letter ᆪ (U+11AA) is not decomposed any further in UAX #15, although it could be decomposed into two simple letters ᆨ (U+11A8) and ᆺ (U+11BA).

At step D1, an original syllable (U+AC03) is decomposed into three letters:

갃 (U+AC03) --> ᄀ + ᅡ + ᆪ (U+1100 1161 11AA)

Then, at step D2, a 2-complex letter (U+11AA) is further decomposed into two simple letters (U+11A8) and (U+11BA):

ᆪ (U+11AA) --> ᆨ (U+11A8) + ᆺ (U+11BA)

Therefore, the final decomposition is:

- An original syllable: 갃 (U+AC03)
  --> [NFD] ᄀ + ᅡ + ᆨ + ᆺ (U+1100 1161 11A8 11BA)

Now let's take some more examples.

Example 3.1

Example 3.1a
ㅗㅐ (U+116B) --> [NFD] ㅗ (U+1169) + ㅏ (U+1161) + ㅣ (U+1175)

Example 3.1b
ㅗㅏ (U+116A) + ㅣ (U+1175) --> [NFD] ㅗ (U+1169) + ㅏ (U+1161) + ㅣ (U+1175)

Example 3.1c
ㅗ (U+1169) + ㅐ (U+1162) --> [NFD] ㅗ (U+1169) + ㅏ (U+1161) + ㅣ (U+1175)

Example 3.1d
ㅗ (U+1169) + ㅏ (U+1161) + ㅣ (U+1175) --> (Since all letters are simple, we cannot decompose any further.)

The above four original sequences are considered equal when rendered. However, UAX #15 decomposition canNOT handle any of the four examples above, since that process handles only Modern syllables. In contrast, our new decomposition process decomposes a complex letter into simple letters so that four sequences become the same in a decomposed form.

Example 3.2 Our new decomposition process decomposes the following three sequences as follows:

Example 3.2a: 개 (U+AC1C) --> [NFD] ㄱ (U+1100) + ㅏ (U+1161) + ㅣ (U+1175)
Example 3.2b: ㄱ (U+1100) + ㅐ (U+1162) --> [NFD] ㄱ (U+1100) + ㅏ (U+1161) + ㅣ (U+1175)
Example 3.2c: ㄱ (U+1100) + ㅏ (U+1161) + ㅣ (U+1175) --> (no further decomposition, since all letters are already simple.

The above three original sequences are considered equal when rendered and our new decompositions of them are the same.

In contrast, the results according to the UAX #15 canonical decomposition are shown below:
Example 3.2a: 개 (U+AC1C) --> ㄱ (U+1100) + ㅐ (U+1162)
Example 3.2b: ㄱ (U+1100) + ㅐ (U+1162) --> (no further decomposition)
Example 3.2c: ㄱ (U+1100) + ㅏ (U+1161) + ㅣ (U+1175) --> (no further decomposition)

As you can see, in a UAX #15 decomposed form, two sequences of examples 3.2a and 3.2b are equal. However, they are different from the sequence of example 3.2c.
In contrast, our new decomposition further decomposes the complex letter into

simple letters so that the three sequences of examples 3.2a, 3.2b, and 3.2c are all equal.

The above examples 3.1 and 3.2 show why our new canonical decomposition for Hangeul is better the one in UAX #15.

Example 3.3: Now let's consider Old Hangeul syllables and Old complex letters. UAX #15 decomposition does not accept at all Old Hangeul syllables or Old complex letters. In contrast, our new decomposition process handles Old syllables and Old complex letters just like Modern syllables and Modern complex letters.

- original: 봬 (U+1122) + ᅦ (U+1161)
  --> [NFD] ᄇ (U+1107 + ᄉ (U+1109) + ᄀ (U+1100) + ᅦ (U+1161)

3.3 A new canonical composition process for Hangeul

In ISO/IEC 10646-1, 11,172 Modern Hangeul syllables in Wanseong form are included; however, no Old Hangeul syllables in Wanseong form are included. Given a decomposition, we can compose letters into Modern syllables in Wanseong form (U+AC00 - D7A3) easily, but not into Old syllables in Wanseong form.

To be able to apply consistent canonical composition to both Modern and Old syllables, we suggest that letters be composed into complex letters, not into syllables.

In Unicode 3.0 Standard [Unic30] and UAX #15 [UAX15], canonical decomposition of Modern Hangeul syllables is defined; however, the author could not find an "explicit" definition of canonical composition for Hangeul.

There is a description of a composition for Hangeul in the book. Even if we assume that the composition is a canonical composition, the composition is incomplete in the sense that it handles only Modern syllables, but not Old syllables. In other words, the composition algorithm composes letters into Modern syllables; however, we cannot compose a sequence of letters corresponding to an Old Hangeul syllable into an Old syllable in Wanseong form, since there are no Old Hangeul syllables in ISO/IEC 10646.

In this paper, we propose a new canonical composition which can handle both Modern and Old syllables consistently. Given a decomposition composed of simple letters only, we can take the following steps to get a canonical composition.

Step C1) A sequence of simple letters in a canonical decomposition of Hangeul are grouped into syllable-initial, syllable-peak, and possibly syllable-final

letters.

Step C2) For each group of syllable-initial, syllable-peak or syllable-final simple letters, compose as follows:
  - For a simple letter, return the simple letter;
  - For a group of two simple letters composed of letter-1 and letter-2:
    If there is a 2-complex letter, letter-1-2, composed of letter-1 and letter-2, then return letter-1-2;
    otherwise return a sequence of letter-1 and letter-2;
  - For a group of three simple letters composed of letter-1, letter-2, and letter-3:
    If there is a 3-complex letter, letter-1-2-3, composed of letter-1, letter-2, and letter-3, then return letter-1-2-3;
    else if there is a 2-complex letter, letter-1-2, composed of letter-1 and letter-2, then return a sequence of letter-1-2 and letter-3;
    else if there is a 2-complex letter, letter-2-3, composed of letter-2 and letter-3, then return a sequence of letter-1 and letter-2-3;
    otherwise return a sequence of letter-1, letter-2, and letter-3;

Note. It is assumed that a complex letter has at most three simple letters, which is true for any Hangeul syllable/letter found up to date. However, the above algorithm can be easily extended to accommodate more than three simple letters in a group.

  First, we will review the example in [UAX15]:
  - An original syllable: 갃 (U+AC03)
    --> [NFD] ㄱ + ㅏ + ㄳ (U+1100 1161 11AA) --> [NFC] 갃 (U+AC03)

  In contrast, our new composition composes into complex letters, not into syllables:
  - an original syllable: 갃 (U+AC03)
    --> [NFD] ㄱ + ㅏ + ㄱ + ㅅ (U+1100 1161 11A8 11BA)
    --> [NFC] ㄱ + ㅏ + ㄳ (U+1100 1161 11AA)

  Now let's continue with examples 3.1, 3.2, and 3.3 above. All four original sequences in example 3.1 ended up with the same decomposition as shown below:

  [NFD] ㅗ (U+1169) + ㅏ (U+1161) + ㅣ (U+1175)

Its composition is shown below:

  --> [NFC] ㅙ (U+116B)

  Likewise, all three sequences in example 3.2 above ended up with the same decomposition as shown below:

[NFD] ㄱ (U+1100) + ㅏ (U+1161) + ㅣ (U+1175)

Its composition is shown below:

--> [NFC] ㄱ (U+1100) + ㅐ (U+1162)

Now let's consider Old Hangeul syllables and Old complex letters. UAX #15 composition does not accept at all Old Hangeul syllables or Old complex letters. In contrast, our new composition process handles Old syllables and Old complex letters just like Modern syllables and Modern complex letters.

Canonical decomposition in example 3.3 above was:

[NFD] ㅂ (U+1107 + ㅅ (U+1109) + ㄱ (U+1100) + ㅏ (U+1161).

Its composition is shown below:

--> [NFC] ᄡ (U+1122) + ㅏ (U+1161)

3.4 A canonical composition process for Hangeul/option-Modern-syl

Since we frequently use Modern Hangeul syllables, we could modify the proposed composition process slightly to utilize Modern Hangeul syllables in Wanseong form included in ISO/IEC 10646. A canonical composition modified as shown in this section will be called "composition process/option-Modern-syl".

If, after composition as described in section 3.3, a sequence of letters is found to be a Modern syllable, we can further compose the sequence of letters into a Modern syllable.

Although the author believes that the composition in section 3.3 is better than the one in this section, sometimes we may find this modified composition process useful.

This modified composition process is basically as follows:

a) For Old Hangeul, we use our new composition process; that is, we compose simple letters into complex letters, if possible.

b) For Modern Hangeul, we use the composition process described in UAX #15; that is, after applying the new composition process, we further compose letters into a Modern syllable, if possible.

4. Conclusions

In this paper, the author proposed new canonical decomposition process (NFD: Normalization Form D) and composition processes (NFC: Normalization Form C) for Hangeul, Korean script. The new processes can handle both Modern and Old Hangeul in a consistent way. Therefore, the new processes are more general than the ones described in UAX #15 [UAX15], which handle only Modern Hangeul syllables.

Since NFD and NFC will play an important role in sorting and searching, Internet-related applications such as Multilingual Domain names, databases, etc., we need to establish good NFD and NFC for Hangeul before they are widely used.

The author hopes that NFD and NFC proposed in this paper will be reviewed and reflected in ISO/IEC 14651 International String Ordering and Comparison [ISO14651], Unicode Standard Annex #15 [UAX15], Nameprep for IDN (Internationalized Domain Name) [Nameprep], and other related documents.

References.

[ISO10646] ISO/IEC 10646-1:2000, Information Technology -- Universal Multiple-Octet Coded Character Set (UCS) -- Part 1: Architecture and Basic Multilingual Plane. 2000-10-05.

[ISO11941] ISO TS (Technical Specification) 11941, Information and Documentation -- Transliteration of Korean script into Latin characters. First edition, 1996-12-01.

[ISO14651] ISO/IEC 14651, International String Ordering and Comparison - Method for comparing character strings and description of the common template tailorable ordering

[KimK99] A story about Hangeul inside computers, Vol. 2. KIM, Kyongsok. Busan National University Press. Dec. 31, 1999 [in Hangeul (Korean)].

[Nameprep] Paul Hoffman, Marc Blanchet, "Preparation of Internationalized Host Names", Feb 2001, http://www.ietf.org/internet-drafts/draft-ietf-idn-nameprep-03.txt

[UAX15] Unicode Standard Annex #15. Technical Report. Unicode Normalization Forms. Mark Davis and Martin Duerst. 2001-03-23.

[Unic30] The Unicode Standard Version 3.0, The Unicode Consortium. 2000.

Appendices:

Appendix 1. Decomposition of complex letters into simple letters.

1) Decomposition of syllable-initial complex letters into simple letters.

U+1101 -> U+1100 1100
U+1104 -> U+1103 1103
U+1108 -> U+1107 1107
U+110A -> U+1109 1109
U+110D -> U+110C 110C
U+1113 -> U+1102 1100
U+1114 -> U+1102 1102
U+1115 -> U+1102 1103
U+1116 -> U+1102 1107
U+1117 -> U+1103 1100
U+1118 -> U+1105 1102
U+1119 -> U+1105 1105
U+111A -> U+1105 1112
U+111B -> U+1105 110B
U+111C -> U+1106 1107
U+111D -> U+1106 110B
U+111E -> U+1107 1100
U+111F -> U+1107 1102

U+1120 -> U+1107 1103
U+1121 -> U+1107 1109
U+1122 -> U+1107 1109 1100
U+1123 -> U+1107 1109 1103
U+1124 -> U+1107 1109 1107
U+1125 -> U+1107 1109 1109
U+1126 -> U+1107 1109 110C
U+1127 -> U+1107 110C
U+1128 -> U+1107 110E
U+1129 -> U+1107 1110
U+112A -> U+1107 1111
U+112B -> U+1107 110B
U+112C -> U+1107 1107 110B
U+112D -> U+1109 1100
U+112E -> U+1109 1102
U+112F -> U+1109 1103

U+1130 -> U+1109 1105
U+1131 -> U+1109 1106

U+1132 -> U+1109 1107
U+1133 -> U+1109 1107 1100
U+1134 -> U+1109 1109 1109
U+1135 -> U+1109 110B
U+1136 -> U+1109 110C
U+1137 -> U+1109 110E
U+1138 -> U+1109 110F
U+1139 -> U+1109 1110
U+113A -> U+1109 1111
U+113B -> U+1109 1112
U+113D -> U+113C 113C
U+113F -> U+113E 113E

U+1140 -> U+1107 1103
U+1141 -> U+110B 1100
U+1142 -> U+110B 1103
U+1143 -> U+110B 1106
U+1144 -> U+110B 1107
U+1145 -> U+110B 1109
U+1146 -> U+110B 1140
U+1147 -> U+110B 110B
U+1148 -> U+110B 110C
U+1149 -> U+110B 110E
U+114A -> U+110B 1110
U+114B -> U+110B 1111
U+114D -> U+110C 110B
U+114F -> U+114E 114E

U+1151 -> U+1150 1150
U+1152 -> U+110E 110F
U+1153 -> U+1103 1112
U+1156 -> U+1111 1107
U+1157 -> U+1111 110B
U+1158 -> U+1112 1112

2) Decomposition of syllable-peak complex letters into simple letters.

U+1162 -> U+1161 1175
U+1164 -> U+1163 1175
U+1166 -> U+1165 1175
U+1168 -> U+1167 1175
U+116A -> U+1169 116A
U+116B -> U+1169 116A 1175

U+116C -> U+1169 1175
U+116F -> U+116E 1165
U+1170 -> U+116E 1165 1175
U+1171 -> U+116E 1175
U+1174 -> U+1173 1175
U+1176 -> U+1161 1169
U+1177 -> U+1161 116E
U+1178 -> U+1163 1169
U+1179 -> U+1163 116D
U+117A -> U+1165 1169
U+117B -> U+1165 116E
U+117C -> U+1165 1173
U+117D -> U+1167 1169
U+117E -> U+1167 116E
U+117F -> U+1169 1165
U+1180 -> U+1169 1165 1175
U+1181 -> U+1169 1167 1175
U+1182 -> U+1169 1169
U+1183 -> U+1169 116E
U+1184 -> U+116D 1163
U+1185 -> U+116D 1163 1175
U+1186 -> U+116D 1167
U+1187 -> U+116D 1169
U+1188 -> U+116D 1175
U+1189 -> U+116E 1161
U+118A -> U+116E 1161 1175
U+118B -> U+116E 1165 1173
U+118C -> U+116E 1167 1175
U+118C -> U+116E 1167 1175
U+118D -> U+116E 116E
U+118F -> U+1172 1165
U+1190 -> U+1172 1165 1175
U+1191 -> U+1172 1167
U+1192 -> U+1172 1167 1175
U+1193 -> U+1172 116E
U+1194 -> U+1172 1175
U+1195 -> U+1173 116E
U+1196 -> U+1173 1173
U+1197 -> U+1173 1175 116E
U+1198 -> U+1175 1161
U+1199 -> U+1175 1163
U+119A -> U+1175 1169
U+119B -> U+1175 116E
U+119C -> U+1175 1173

U+119D -> U+1175 119E
U+119F -> U+119E 1165
U+11A0 -> U+119E 116E
U+11A1 -> U+119E 1175
U+11A2 -> U+119E 119E

3) Decomposition of syllable-final complex letters into simple letters.

U+11A9 -> U+11A8 11A8
U+11AA -> U+11A8 11BA
U+11AC -> U+11AB 11BD
U+11AD -> U+11AB 11C2

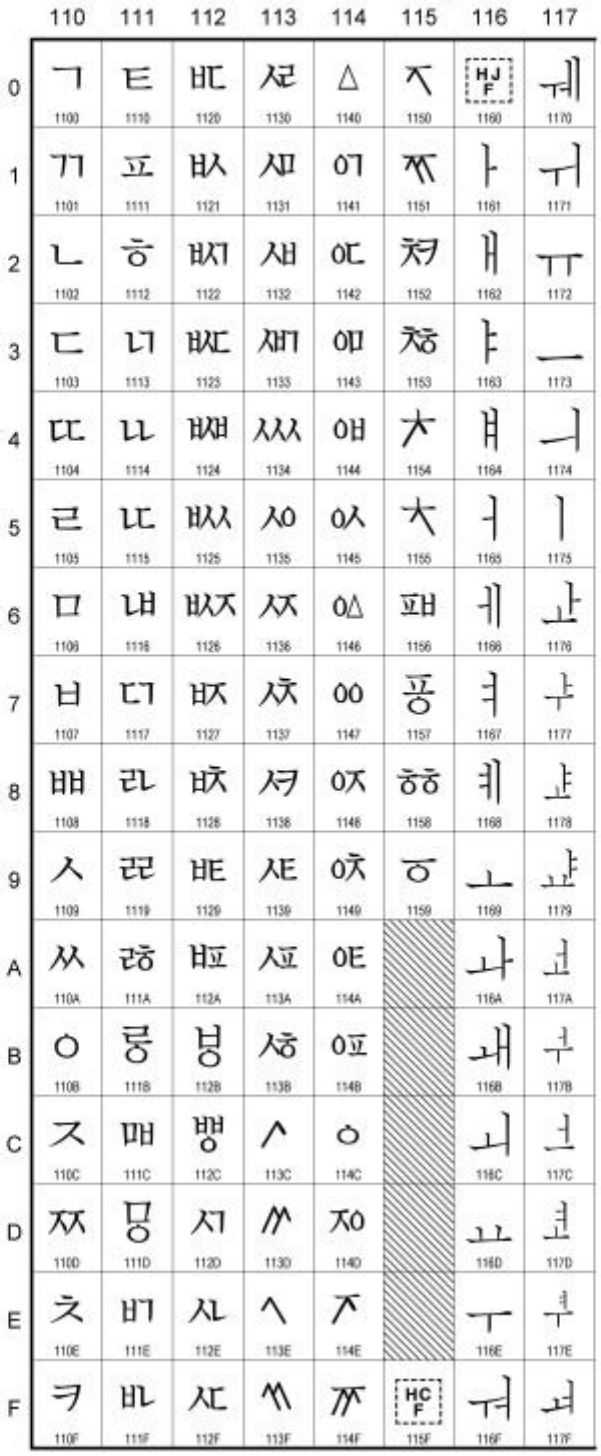
U+11B0 -> U+11AF 11A8
U+11B1 -> U+11AF 11B7
U+11B2 -> U+11AF 11B8
U+11B3 -> U+11AF 11BA
U+11B4 -> U+11AF 11C0
U+11B5 -> U+11AF 11C1
U+11B6 -> U+11AF 11C2
U+11B9 -> U+11B8 11BA
U+11BB -> U+11BA 11BA
U+11C3 -> U+11A8 11AF
U+11C4 -> U+11A8 11BA 11A8
U+11C5 -> U+11AB 11A8
U+11C6 -> U+11AB 11AE
U+11C7 -> U+11AB 11BA
U+11C8 -> U+11AB 11EB
U+11C9 -> U+11AB 11C0
U+11CA -> U+11AE 11A8
U+11CB -> U+11AE 11AF
U+11CC -> U+11AF 11A8 11BA
U+11CD -> U+11AF 11AB
U+11CE -> U+11AF 11AE
U+11CF -> U+11AF 11AE 11C2
U+11D0 -> U+11AF 11AF
U+11D1 -> U+11AF 11B7 11A8
U+11D2 -> U+11AF 11B7 11BA
U+11D3 -> U+11AF 11B8 11BA
U+11D4 -> U+11AF 11B8 11C2
U+11D5 -> U+11AF 11B8 11BC
U+11D6 -> U+11AF 11BA 11BA
U+11D7 -> U+11AF 11EB

U+11D8 -> U+11AF 11BF
U+11D9 -> U+11AF 11F9
U+11DA -> U+11B7 11A8
U+11DB -> U+11B7 11AF
U+11DC -> U+11B7 11B8
U+11DD -> U+11B7 11BA
U+11DE -> U+11B7 11BA 11BA
U+11DF -> U+11B7 11EB
U+11E0 -> U+11B7 11BE
U+11E1 -> U+11B7 11C2
U+11E2 -> U+11B7 11BC
U+11E3 -> U+11B8 11AF
U+11E4 -> U+11B8 11C1
U+11E5 -> U+11B8 11C2
U+11E6 -> U+11B8 11BC
U+11E7 -> U+11BA 11A8
U+11E8 -> U+11BA 11AE
U+11E9 -> U+11BA 11AF
U+11EA -> U+11BA 11B8
U+11EC -> U+11BC 11A8
U+11ED -> U+11BC 11A8 11A8
U+11EE -> U+11BC 11BC
U+11EF -> U+11BC 11BF
U+11F1 -> U+11F0 11BA
U+11F2 -> U+11F0 11EB
U+11F3 -> U+11C1 11B8
U+11F4 -> U+11C1 11BC
U+11F5 -> U+11C2 11AB
U+11F6 -> U+11C2 11AF
U+11F7 -> U+11C2 11B7
U+11F8 -> U+11C2 11B8
* * *

Appendix 2. Hangeul letters: U+1100 - 117F [KS X 1005-1] (contiuned)

표 35 – 행 11: 한글 자모

Table 35 - Row 11: Hangul Jamo



	110	111	112	113	114	115	116	117
0	ᄀ 1100	ᄐ 1110	ᄠ 1120	ᄰ 1130	ᅀ 1140	ᅐ 1150	HJF 1160	ᅿ 1170
1	ᄁ 1101	ᄑ 1111	ᄡ 1121	ᄱ 1131	ᅁ 1141	ᅑ 1151	ᆁ 1161	ᆀ 1171
2	ᄂ 1102	ᄒ 1112	ᄢ 1122	ᄲ 1132	ᅂ 1142	ᅒ 1152	ᆂ 1162	ᆄ 1172
3	ᄃ 1103	ᄓ 1113	ᄣ 1123	ᄳ 1133	ᅃ 1143	ᅓ 1153	ᆃ 1163	ᆅ 1173
4	ᄄ 1104	ᄔ 1114	ᄤ 1124	ᄴ 1134	ᅄ 1144	ᅔ 1154	ᆄ 1164	ᆆ 1174
5	ᄅ 1105	ᄕ 1115	ᄥ 1125	ᄵ 1135	ᅅ 1145	ᅕ 1155	ᆅ 1165	ᆇ 1175
6	ᄆ 1106	ᄖ 1116	ᄦ 1126	ᄶ 1136	ᅆ 1146	ᅖ 1156	ᆆ 1166	ᆈ 1176
7	ᄇ 1107	ᄗ 1117	ᄧ 1127	ᄷ 1137	ᅇ 1147	ᅗ 1157	ᆇ 1167	ᆉ 1177
8	ᄈ 1108	ᄘ 1118	ᄨ 1128	ᄸ 1138	ᅈ 1148	ᅘ 1158	ᆈ 1168	ᆊ 1178
9	ᄉ 1109	ᄙ 1119	ᄩ 1129	ᄹ 1139	ᅉ 1149	ᅙ 1159	ᆉ 1169	ᆋ 1179
A	ᄊ 110A	ᄚ 111A	ᄪ 112A	ᄺ 113A	ᅊ 114A		ᆊ 116A	ᆌ 117A
B	ᄋ 110B	ᄛ 111B	ᄫ 112B	ᄻ 113B	ᅋ 114B		ᆋ 116B	ᆍ 117B
C	ᄌ 110C	ᄜ 111C	ᄬ 112C	ᄼ 113C	ᅌ 114C		ᆌ 116C	ᆎ 117C
D	ᄍ 110D	ᄝ 111D	ᄭ 112D	ᄽ 113D	ᅍ 114D		ᆍ 116D	ᆏ 117D
E	ᄎ 110E	ᄞ 111E	ᄮ 112E	ᄾ 113E	ᅎ 114E		ᆎ 116E	ᆐ 117E
F	ᄏ 110F	ᄟ 111F	ᄯ 112F	ᄿ 113F	ᅏ 114F	HCF 115F	ᆏ 116F	ᆑ 117F

G = 00
P = 00

Hangeul letters: U+1180 - 11FF [KS X 1005-1]

표 36 - 행 11: 한글 자모

Table 36 - Row 11: Hangul Jamo

G = 00
P = 00

Appendix 3. Hangeul syllables: U+AC00 - ACAF [KS X 1005-1]

(Out of 11,172 Hangeul syllables, only 176 syllables are shown below.)

표 98 - 행 AC: 한글 글자마디

Table 98 - Row AC: Hangul Syllables

	AC0	AC1	AC2	AC3	AC4	AC5	AC6	AC7	AC8	AC9	ACA
0	가 AC00	감 AC10	갠 AC20	갰 AC30	걀 AC40	걬 AC50	걠 AC60	거 AC70	검 AC80	겐 AC90	겠 ACA0
1	각 AC01	갑 AC11	갡 AC21	갱 AC31	걁 AC41	걑 AC51	걡 AC61	걱 AC71	겁 AC81	겑 AC91	겡 ACA1
2	갂 AC02	값 AC12	갢 AC22	갲 AC32	걂 AC42	걒 AC52	걢 AC62	걲 AC72	겂 AC82	겒 AC92	겢 ACA2
3	갃 AC03	갓 AC13	갣 AC23	갳 AC33	걃 AC43	걓 AC53	걣 AC63	걳 AC73	것 AC83	겓 AC93	겣 ACA3
4	간 AC04	갔 AC14	갤 AC24	갴 AC34	걄 AC44	개 AC54	갬 AC64	건 AC74	걷 AC84	겔 AC94	겤 ACA4
5	갅 AC05	강 AC15	갥 AC25	갵 AC35	걅 AC45	걕 AC55	걥 AC65	걵 AC75	걹 AC85	겕 AC95	겥 ACA5
6	갆 AC06	갖 AC16	갦 AC26	갶 AC36	걆 AC46	걖 AC56	걦 AC66	걶 AC76	걺 AC86	겖 AC96	겦 ACA6
7	갇 AC07	갗 AC17	갧 AC27	갷 AC37	걇 AC47	걗 AC57	걧 AC67	걷 AC77	걻 AC87	겗 AC97	겧 ACA7
8	갈 AC08	갘 AC18	갨 AC28	갸 AC38	걈 AC48	걘 AC58	걨 AC68	걸 AC78	걼 AC88	겘 AC98	겨 ACA8
9	갉 AC09	갚 AC19	갩 AC29	갹 AC39	걉 AC49	걙 AC59	걩 AC69	걹 AC79	걽 AC89	겙 AC99	격 ACA9
A	갊 AC0A	갛 AC1A	갪 AC2A	갺 AC3A	걊 AC4A	걚 AC5A	걪 AC6A	걺 AC7A	겁 AC8A	겚 AC9A	겪 ACAA
B	갋 AC0B	개 AC1B	갫 AC2B	갻 AC3B	걋 AC4B	걛 AC5B	걫 AC6B	걻 AC7B	겋 AC8B	겛 AC9B	겫 ACAB
C	갌 AC0C	개 AC1C	갬 AC2C	갼 AC3C	걌 AC4C	걜 AC5C	걬 AC6C	걼 AC7C	게 AC8C	겜 AC9C	견 ACAC
D	갍 AC0D	객 AC1D	갭 AC2D	갽 AC3D	걍 AC4D	걝 AC5D	걭 AC6D	걽 AC7D	겍 AC8D	겝 AC9D	겭 ACAD
E	갎 AC0E	갞 AC1E	갮 AC2E	갾 AC3E	걎 AC4E	걞 AC5E	걮 AC6E	걾 AC7E	겎 AC8E	겞 AC9E	겮 ACAE
F	갏 AC0F	갟 AC1F	갯 AC2F	갿 AC3F	걏 AC4F	걟 AC5F	걯 AC6F	걿 AC7F	겏 AC8F	겟 AC9F	겯 ACAF

G = 00
P = 00

Author's Vita:

KIM, Kyongsok (GIM, Gyeongseog) is a Professor in Division of Computer Science and Engineering at Busan National University, Busan 609-735, South Korea. His e-mail address is gimgs@HANGEUL.pnu.edu. He received Bachelor's and Master's degrees at Seoul National University, South Korea, and Ph. D in Computer Science at University of Illinois at Urbana-Champaign, USA in 1988.

He worked at North Dakota State University as an assistant professor from 1988 to 1992. He has been working at Busan National University since 1992. He stayed in Dept. of Phonetics and Linguistics at University College London, U. K from 1997 to 1998 as a visiting professor.

Currently he is a chairperson of Korea JTC1/SC2 committee on Code; a member of Korea TC46 committee on Document Information and Hangeul Romanization; a member of Korea SC22 (Programming Languages, esp. WG20 - Internationalization); a member of Name Committe at KRNIC (Korea Network Information Center). He is a member of KLS (Korean Language Society), KLIS (Korean Language Information Society), and KISS (Korea Information Science Society).

Information about Graphics Files for Figures:

There are four graphics files as follows:

- app2a.jpg for the first figure in appendix 2
- app2b.jpg for the second figure in appendix 2
- app3.jpg for the figure in appendix 3
- gimgs12w.jpg for author's photo

* * *