Abstract

This document describes code point reordering to improve ACE compression algorithms. Being based on character frequency and character adjacency statistics for major character sets, this reordering can be easily implemented only with simple character mapping tables without adding complexity to existing ACE algorithms.

When applied to DUDE and AMC-ACE-W, this reordering greatly improves both ACEs’ compression ratios for Hangul, Chinese, Vietnamese, Katakana and European domains. Interestingly, reordered DUDE shows better or equal compression ratio than both bare AMC-ACE-W and reordered AMC-ACE-W.
Differences from version 0.9

version 1.0 differs from version 0.9 in four respects:

1) For Hangul, it does not use Hangul jamo frequency order any more:
   instead, it adopts new reordering based on both hangul character frequency and adjacency in words
   used as business names in Korea.

2) For Unified Han, like for Hangul, it adopts new reordering table that reflects
   han character frequency and adjacency in words used as business names in China/Taiwan/Japan.

3) new supports for Japanese katakana

4) additional implementation of reordering with AMC-ACE-W

Overview

Pursuing shorter ACE labels is justified to save memory resources and to reduce internet traffic even for domains of average length in various application/core internet protocols.

Both 11172 Hangul syllables and 24000 or more CJK Han syllables occupy roughly half of the entire unicode space. Their lexicographical ordering (not in frequency ordering) makes various ACE compression technique work poorly for them, because they are spread evenly throughout those wide ranges.

The most frequent 256 Hangul syllables has cumulative frequency sum of 88.2% and for the case of top 512 ones, it reaches 99.9%.

The most frequent 256 Han letters has cumulative frequency sum of 58.2% and for the cases of top 512, 1024, 2048 and 4096 ones, it reaches 72.8, 85.9, 95.4 and 99.4%, respectively.

Even, Latin characters code range, including ‘a’ - ‘z’ has lexicographical order that does not reflect the fact that ‘s’, ‘t’ and ‘r’ are more frequently used than ‘j’, ‘k’ and ‘h’.

Most ACE algorithms show good compression ratio when frequently used characters are located in narrow code ranges. Especially, to reduce DUDE XOR distance, we can make the narrow area fit in aligned blocks of 16,256 or 4096 code points.

Unified Han and Hangul

Most frequently used 4096 Traditional Chinese/Simplified Chinese/Japanese Kanji letters are reordered into single aligned block of 4096 code points. Their combinations are estimated to form
almost 99% of modern Chinese business names.

Most frequently used 888 Hangeul syllables are reordered into the lower portion of single aligned block of 4096 code points. Their combinations are expected to cover almost 99% of modern Hangul business names.

In fact, every block of 256 code points in these reordered areas is designed to reflect not only character frequency order, but also to reflect adjacency preference derived from statistics on major business category names or famous regional names in Eastern Asia.

For example, there is a frequent Korean industrial category name ‘jeon-ja’ (electronics). In pure frequent-oriented ordering, its two component Hangul syllables ‘jeon’ and ‘ja’ should have been put far apart from each other. But in this new adjacency-adjusted frequency ordering, they are put together in a single row (u+??0 ~ u+??f) in order to reduce the XOR distance toward single quintet (for DUDE).

The Hangul frequency mapping tables and its statistical data are constructed from business names found in internet directory sites at {cn|tw|kr}.yahoo.com.

Japanese Katakana and Hiragana

Japanese hiragana (u+3040 ~ u+309f) shows relatively even frequency distribution in Japanese business names. And it is often replaced with its Kanji (Japanese Han letter) equivalent in registered business names. I have no mapping table for hiragana, yet.

Japanese katakana (u+30a0 ~ u+30ff) has been widely used to express foreign or English words in Japanese. Most frequent 10 katakanas’ cumulative frequency is estimated to be around 40%. The katakana frequency mapping table is constructed from business names found in internet directories sites such as www.yahoo.co.jp.

Basic Latin

Basic Latin row u+0070 ~ u+007f has ‘p’, ‘r’, ‘s’, ‘t’ and ‘u’ which are more frequently used in European nouns than ‘a’, ‘j’, ‘k’, ‘f’ and ‘g’ in u+0060 ~ u+006f row which includes most frequently used ‘a’~‘o’.

If these two sets of 5 characters are swapped character-wise, ‘p’, ‘r’, ‘s’, ‘t’, ‘u’ go into the u+0060 ~ u+006f row.

Any character sequence only from this single aligned block of 16 codes has XOR-distance or code window length shorter than 0x10 and makes DUDE and other ACEs do good compression.

Extended Latin and Combining Diacritical Marks

First 6 rows from Latin Extension A (u+0100 ~ u+015f) and 6 rows from Basic Latin & Latin-1 Supplement (u+0000 ~ u+002f and u+0080 ~ u+00a0) are swapped.
First 3 rows from Combining Diacritical Marks (u+0300 ~ u+032f) and
3 rows from Latin-1 Supplement (u+00B0 ~ u+00df) are also swapped.

This makes frequently used parts of Latin Extended-A and
Combining Diacritical Marks go into first align block of 256 codes
points (u+0000 ~ u+00ff). Any character sequences from this single
block make XOR-distance or code window length much shorter
than 0x100.

This improvement benefits especially East-European and Vietnamese
that use Latin Extended A and Combining Diacritical Marks.

Other character sets

For Arabic, Cyrillic and Hindi etc, we can devise similar
frequency mapping tables as that for katakana.

Modified Encoding procedure of DUDE implementation of this idea

All ordering of nybbles and quintets is big-endian (most significant
first). A nybble is 4 bits. XOR is bitwise exclusive or.

This modification is hyphen-safe.
Hyphen encoding and decoding are not affected by this modification.

let prev = 96
for each input integer n (in order) do begin
  if n == 45 then output hyphen minus
  else begin
    n = reorder(n) // ******** ADDED *********
    let diff = prev XOR n
    extract the least significant nybbles of diff, as few as are
    sufficient to hold all the nonzero bits (but at least one)
    prepend 0 to the last nybble and 1 to the rest
    output base-32 characters corresponding to the quintets
    let prev = n
  end
end

The encoder must either correctly handle all integer values that can
be represented in the type of its input, or it must check whether
the input contains values that it cannot handle and return an error
if so. Under no circumstances may it produce incorrect output.

Modified Decoding procedure of DUDE implementation of this idea

let prev = 96
while the input string is not exhausted do begin
  if the next character is hyphen-minus then output 45
  else begin
    input characters and convert them to quintets until
    encountering a quintet beginning with 0
    fail upon encountering a non-base-32 character or end-of-input
    strip the first bit of each quintet

concatenate the resulting nybbles to form diff
let prev = prev XOR diff

output restore_order(prev)  // ******** MODIFIED **********

end
end
encode the output sequence and compare it to the input string
fail if they are not equal

Modified Encoding and decoding algorithms of AMC-ACE-W for this idea

(This modification does not affect literal mode of AMC-ACE-W).

procedure initialize(refpoint,style,literal):
  let refpoint[1..5] = (0xE0, 0xA0, 0, 0, 0x10000)
  let style = 0
  let literal = false

procedure update(refpoint,style,n,k):
  # Update the active style and reference points based on
  # the latest code point (n) and the number of base-32
  # characters used to represent it (k).
  let style = k < 3 ? 0 : k > 3 ? 1 : style
  let refpoint[1] = (n >> 4) << 4
  if (k > 2) then let refpoint[2] =
    n is in 00A0..017F ? 0xA0 : (n >> 8) << 8
  if (k > 3) then let refpoint[3] = n is in 3000..9FFF ? 0x4E00 :
    style == 1 and n is in 0xA000..0xD7FF ? 0x8800 : (n >> 12) << 12

procedure encode:
  constant maxdelta[0][1..5] = (0xF, 0xFF, 0xFFF, 0xFFFF, 0xFFFFF)
  constant maxdelta[1][2..5] = (     0xFF, 0x4FFF, 0xFFFF, 0xFFFFF)
  initialize(refpoint,style,literal)
  for each input code point n (in order) do begin
    # Check code point range to avoid array bounds errors later:
    if n is not in 0..10FFFF then fail
    if n == 0x2D then output two hyphen-minuses
    else if n represents an LDH character then begin
      # Letter/digit is encoded literally, so get into literal mode.
      if not literal then output hyphen-minus
      let literal = true
      output the character represented by n
    end
    else begin
      # Non-LDH code point is encoded in base-32.
      # Compute the number of base-32 characters to use:
      n = reorder(n)  // ADDED **********************
      for k = 1 + style to infinity do begin
        let delta = n - refpoint[k]
        if delta is in 0..maxdelta[style][k] then break
      end
      # Switch to base-32 mode if necessary:
      if literal then output hyphen-minus
      let literal = false
      # Check for the extended delta of style 1 window 3:
      if k == 3 and delta >= 0x1000
      then represent (delta - 0x1000) in base 32 as three quintets
      else begin
# Normal case, four bits per quintet:
represent delta in base 16 as k quartets
prepend 0 to the last quartet and 1 to each of the others
output a base-32 character corresponding to each quintet
update(refpoint,style,n,k)
end

procedure decode:
initialize(refpoint,style,literal)
while the input string is not exhausted do begin
read the next character into c
# Unpaired hyphen-minus toggles the mode:
if c is hyphen-minus and the next character is not
then read the next character into c and toggle literal
# Double hyphen-minus represents 0x2D:
if c is hyphen-minus
then read the next character and append 0x2D to history
else if literal then append the code point of c to history
else begin
# Decode a base-32 sequence.
convert c to a quintet
while a quintet beginning with 0 has not been seen
do read and convert up to four more characters
concatenate the lowest four bits of each quintet to form delta
# Check for the extended delta of style 1 window 3:
if style == 1 and there was only one quintet then begin
read two characters and convert them to two more quintets
concatenate delta and the two quintets to form a new delta
let delta = delta + 0x1000
end
let k = the number of quintets decoded
let n = refpoint[k] + delta
update(n,k)
output restore_order(n)  // MODIFIED ***************
end
end
# Enforce the uniqueness of the encoding:
encode the output sequence and compare it to the input string
fail if they are not equal

Example strings
About 30%-58% improvement in DUDE compression ratio is achieved in these Hangul examples.

LDUDE and LAMCW denote reordering-applied DUDE-02 and AMC-ACE-W, respectively. (AMCW for AMC-ACE-W).
Most examples show LDUDE outperforms LAMCW.

(K1) Korean String 1: ( 24 hangul syllables )
u+C138 u+C758 u+BAA8 u+B4E0 u+C0AC u+B78C u+B4E4
u+C774 u+D55C u+AD6D u+C5B4 u+B97C u+C774 u+D574 u+D55C
u+B2E4 u+BA74 u+C5BC u+B9C8 u+B098 u+C88B
(K2) Korean String 2:  ( 9 hangul syllables )
<KRNIC in korean>
U+D55C U+AD6D U+C778 U+D130 U+B137 U+C815 U+BCF4 U+C13C U+D130

DUDE-02 : 7xvItuuNzx5PzsyPz85N97Nz9zA
( 27 chars )
LDUDE : 53C4F5Q7PtwRtMK
( 16 chars, 40% shorter )
AMCW : 7xxIM5wGyjKxeJa2G8ePfw
( 22 chars )
LAMCW : 5s9CwH8JvE5tzMyAK
( 17 chars )

(K4) Korean String 4:  ( 7 hangul syllables )
<Hynics Semiconductor in korean>
U+D558 U+C774 U+B2C9 U+C2A4 U+BC18 U+B3C4 U+CCB4

DUDE-02 : 7yvIz48Fy4sJzxyPzyuJts3Jy3zBy3yPz6Ny8zPz56At7EtswN
( 50 chars )
LDUDE : 5s7NB4EDvHfxDv5Kv6NtIt4R5GwK
( 50 chars )

(K5) Korean String 5:  ( 13 hangul syllables )
U+D558 U+ACBD U+C6B4 U+B3D9 U+C5F0 U+D569 U+BC18 U+D575 U+D2B9 U+BCC4 U+C704 U+C6D0 U+D68C

DUDE-02 : 7yvIz48Fy4sJzxyPzyuJts3Jy3zBy3yPz6Ny8zPz56At7EtswN
( 50 chars )
LDUDE : 5s7NB4EDvHfxDv5Kv6NtIt4R5GwK
( 50 chars )
About 35%~50% improvement in DUDE compression ratio is achieved in these UniHan examples.

(TC1) Traditional Chinese String 1: (16 letters)

DUDE-02: xvve6u3d6t4c87cts

LDUDE: xs8qy7ny9jhyi6f6bb8h-4iy7nyxxbed

AMCW: xvxen8hu8yafz2mc5pcipw7jh7u--xxen8h8ciqcsvynx9i

LAMCW: xvs2q2xcu4m4n6esb6abug--2q2xcusijpq

(TC2) Traditional Chinese String 2: (21 letters)

DUDE-02: xvzht75mts4q694jtwq9qzt

LDUDE: xtwicfz6b99a38q27c2vdd8cz7mzuqdt6izuiy6iaz5nz5zy6by6ib

AMCW: xvths4naacn7mj9fh6vekuvh6ve89vynx9iapbn7m87uyb2v

LAMCW: xtuiukr28q5tgu9i4ukutjk9i3uduspqvyg28quug33kuur28quugh

(TC3) Traditional Chinese String 3: (18 letters)

DUDE-02: z3vq9y8n9usa8w5itz4b6tzgt95iu77hu77h87c4sbovy8xkuxuj87

LDUDE: xwsix5e9kzyq28fhh2phtvgxbwuah8qbtwmyg

AMCW: z3xqnpuh7uq2knfmt7pyyfh7uyyfyzstg4nyu7ayzmbspi75ys

LAMCW: xwyiu7nug3iwu4pkmug4mnv3ky2mu4mnwcdvsiyq

(SC1) Simplified Chinese String 1: (16 letters)

<ministry of foreign trade and economic cooperation, PRC>

DUDE-02: w8wpt7ydt79eu4mvre7yak9puzb7seu8r7wuq85nt27ntv2b3vgt
(SC2) Simplified Chinese String 2: (18 letters)

u+4e2d u+56fd u+4eba u+6c11 u+5927 u+5b66 u+4e2d u+56fd
u+8d22 u+653f u+91d1 u+878d u+653f u+7b56 u+7814 u+7a76
u+4e2d u+5fc3

DUDE-02 : w8wpt27at2whuu4mvxvgwbtwxwmt27a757r82tp9w8qtyxn8u5ct\
  8yjwcvucyvwxtmth8q (69 chars)
LDUDE : xswjf5gu7fu6rb4ifz8dx6ju8gnkwuxwy8fd8rd
  (40 chars, 42.2% shorter)
AMCW : w8up29ps5kdst5uh7ygsup29pm3cb39n8tknbp39hkygwhdysupaqd
  (55 chars)
LAMCW : xsujun3kwwru2abujn36rwsgu8anwsg2auu6f6gjyk
  (41 chars)

About 20%-35% improvement in DUDE compression ratio is achieved in these Japanese Kanji/Katakana examples.

(JP1) Japanese String 1: (25 letters)

U+793E U+56E3 U+6CD5 U+4EBA U+65E5 U+672C U+30CD U+30C3 U+30C8
U+30EF U+30FC U+30AF U+30B9 U+30E1 U+30FC
U+30B7 U+30E7 U+30F3 U+30BB U+30F3 U+30BF U+30FC

DUDE-02 : z3xQu97Pv4vGuuyRu5xRu6Jxz8BQMwHtDxmHuGzNwItPwMxAtE\n  wiWwNwD (60 chars)
LDUDE : xsw8Nu2Cu4RvMGBysxGyCKtHtQCPFtAyPyKtPBGPyAyAyFyR
  (47 chars, 21.6% shorter)
AMCW : z3vQ28DDyxs5KB9fCjnvs6P6DI8R9N4RE9D7F4J8B9N5H8H9D59\n  D5R9 (57 chars)
LAMCW : xs2NwsQu4B3KNPvs6M4JD5E4KIFA5A7P5H4KMPA6A4A6F4K
  (47 chars)

(JP2) Japanese String 2: (15 letters)

U+8CA1 U+56E3 U+6CD5 U+4EBA U+5317 U+6D77 U+9053 U+81EA
U+7136 U+4FDD U+8B77 U+63A8 U+9032 U+5354 U+4F1A

DUDE-02 : 266B74wCv4vGuuyRt74Pv8yA97uEtt5J9s7Nv88Mw4K827R9v3K\n  6vyGt6wQ (60 chars)
LDUDE : xs3Hu9J4RvMt5CFvuGvsRx6Gw5Iz2Ev6Bz1wtJE
  (40 chars, 33.3% shorter)
AMCW : 264B28DDyxsS5KxtH5zNu19kE3yt7PmzBpiTtuxxEtffK
  (46 chars)
LAMCW : xs9HwsQu4B3KvusMvsEytCu4K3uQy8R3Hu2QK
  (40 chars)

(JP3) Japanese String 3: (17 letters)

U+6771 U+4EAC U+90FD U+60C5 U+5831 U+30B5 U+30FC U+30D3 U+30B9
U+7523 U+696D U+5065 U+5EB7 U+4FDD U+967A U+7D44 U+5408

DUDE-02 : yztBu37P78xB9svIv29Eyu22ewJuRyKwx3Kt6wVq3sI87CtyK734\
LDUDE also shows the same good compression ratio for Latin family of scripts.

(L1) Vietnamese: (38 syllables using diacritical marks)
Ta<dotbelow>isaoho<dotbelow>kh<ocirc>ngth<ecirc><hookabove>chi\<hookabove>no<acute><ecirc><dotbelow>t
U+0054 u+0061 u+0323 u+0069 u+0073 u+0061 u+006F u+0068 u+006F u+0323 u+0066 u+006E u+0067 u+0074 u+0066 u+006E u+0309 u+0066 u+0069 u+0309 u+006E u+0066 u+0030 u+0061 u+0074 u+006E u+0066 u+0069 u+006E u+0061 u+0074

DUDE-02: vEvfvwcvwkttcqghhvwnwvid3n3kjtqtn2cv8dvykmbvyavyhbvyqv\yiptpt2dv8mvyrjyBvr2dv6jvwxh (82 chars)
LDUDE: uGuh5c5kckqhh5n4atm3n3ktmdq2cxd7kmb7a7hb7g7irr2dxml7rt\muDvr2dvj5f (66 chars, 19% shorter)

(L2) Spanish: (using basic Latin & Latin Supplement)
Porqu<eacute>nopuedensimplementehablarenEspa<ntilde>n
U+0050 u+006F u+0072 u+0071 u+0075 u+00E9 u+006E u+006F u+0070 u+0075 u+0065 u+0064 u+0065 u+006E u+0073 u+0069 u+006D u+0070 u+006C u+0065 u+006D u+0065 u+006E u+0074 u+0065 u+0068 u+0061 u+0062 u+006C u+0061 u+0072 u+0065 u+006E u+0045 u+0073 u+0070 u+0061 u+00F1 u+006F u+006C

DUDE-02: vAvrtpde3n2hbrttfabbmtptketptnjiimtktbpjdpdtthmuMvgdt\b3a3qd (61 chars)
LDUDE: uAurftmtg2q2hbrcbmfmfcepnjiiimdppdjqpmrmumuMqmb3a3qd
(51 chars, 10 chars (16%) shorter)

(L3) Czech: (using Latin Extended A)
Pro<ccaron>prost<ecaron>nemluvn<iacute>ce<ccaron>eskynelmuvn<iacute>ce
U+0050 u+0072 u+006F u+010D u+0070 u+0072 u+006F u+0073 u+0074 u+011B u+006E u+0065 u+006D u+006C u+0075 u+0076 u+00ED u+010D u+0065 u+0073 u+006F u+0079

DUDE-02: vAuctpyctzpctpntjytzfmtibtd3mt8atyitgtitc (45 chars)
LDUDE: uAukfycypkfezpzmibmtd3m8ayigtik (34 chars, 24% shorter)

Security considerations

ACE-encoded reordered code points are restored in reverse ACE translation and this improvement do not introduce any new security problems into ACE.
LDUDE: Example implementation into DUDE-02

This idea is applicable to any ACEs. LDUDE is a name for DUDE-02 implementation of this idea.

Embedded hangul, han and Latin frequency tables are subject to change with further studies in the next revision of this draft.

In Unix, save this example source code into ldude.c

% cc -o ldude ldude.c
% ./ldude -e < input_file > output_file
% ./ldude -d < output_file

An input file should contain u+????-form code points delimited with spaces or newlines.

/* begin of ldude.c */
/*

/****************************Nguồn: https://www.unicode.org/unicode/standard/

This is ANSI C code (C89) implementing */
/* DUDE (draft-ietf-idn-ldude-01). */
#include <stdio.h>
#include <limits.h>

enum dude_status {
    dude_success,
    dude_bad_input,
    dude_big_output /* Output would exceed the space provided. */
};

enum case_sensitivity { case_sensitive, case_insensitive };

#if UINT_MAX >= 0x1FFFFF
typedef unsigned int u_code_point;
#else
typedef unsigned long u_code_point;
#endif

enum dude_status dude_encode(
    unsigned int input_length,
    const u_code_point input[],
    const unsigned char uppercase_flags[],
    unsigned int *output_size,
    char output[] );

/* dude_encode() converts Unicode to DUDE (without any signature). The input must be represented as an array of Unicode code points (not code units; surrogate pairs are not allowed), and the output will be represented as null-terminated ASCII. The input_length is the number of code points in the input. The output_size is an in/out argument: the caller must pass in the maximum number of characters that may be output (including the terminating null), and on successful return it will contain the number of characters actually output (including the terminating null, so it will be one more than strlen() would return, which is why it is called output_size rather than output_length). The uppercase_flags array must hold input_length boolean values, where nonzero means the corresponding Unicode character should be forced to uppercase after being decoded, and zero means it is caseless or should be forced to lowercase. Alternatively, uppercase_flags may be a null pointer, which is equivalent to all zeros. The encoder always outputs lowercase base-32 characters except when nonzero values of uppercase_flags require otherwise. The return value may be any of the dude_status values defined above; if not dude_success, then output_size and output may contain garbage. On success, the decoder will never need to write an output_size greater than input_length*k+1 if all the input code points are less than 1 << (4*k), because of how the encoding is defined. */

enum dude_status dude_decode(
    enum case_sensitivity case_sensitivity,
    char scratch_space[],
    const char input[],
    unsigned int *output_length,
    u_code_point output[],
    unsigned char uppercase_flags[] );

/* dude_decode() converts DUDE (without any signature) to Unicode. The input must be represented as null-terminated ASCII, and the output will be represented as an array of Unicode code points. The case_sensitivity argument influences the check on the well-formedness of the input string; it must be case_sensitive if case-sensitive comparisons are allowed on encoded strings, case_insensitive otherwise. The scratch_space must point to space at least as large as the input, which will get overwritten (this allows the decoder to avoid calling malloc()). The output_length is an in/out argument: the caller must pass in the maximum number of code points that may be output, and on successful output it will contain the actual number of code points. The uppercase_flags array must have room for at least output_length values, or it may be a null pointer if the case information is not needed. A nonzero flag indicates that the corresponding Unicode character should be forced to uppercase by the caller, while zero means it is caseless or should be forced to lowercase. The return value may be any of the dude_status values defined above; if not dude_success, then output_length, output, and uppercase_flags may contain garbage. On success, the decoder will never need to write an output_length greater than the length of the input (not counting the null terminator), because of how the encoding is defined. */

/***********************************************************/
/* Implementation (would normally go in its own .c file): */

#include <string.h>

/* Character utilities: */

/* base32[q] is the lowercase base-32 character representing the number q from the range 0 to 31. Note that we cannot use string literals for ASCII characters because an ANSI C compiler does not necessarily use ASCII. */

static const char base32[] = {
    97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, /* a-k */
    109, 110, /* m-n */
    112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, /* p-z */
    50, 51, 52, 53, 54, 55, 56, 57, /* 2-9 */
};

/* base32_decode(c) returns the value of a base-32 character, in the range 0 to 31, or the constant base32_invalid if c is not a valid base-32 character. */

enum { base32_invalid = 32 };

static unsigned int base32_decode(char c) {
    if (c < 50) return base32_invalid;
    if (c <= 57) return c - 26;
    if (c < 97) c += 32;
    if (c < 97 || c == 108 || c == 111 || c > 122) return base32_invalid;
    return c - 97 - (c > 108) - (c > 111);
}
/* unequal(case_sensitivity,s1,s2) returns 0 if the strings s1 and s2 */
/* are equal, 1 otherwise. If case_sensitivity is case_insensitive, */
/* then ASCII A-Z are considered equal to a-z respectively. */

static int unequal( enum case_sensitivity case_sensitivity,
        const char s1[], const char s2[] )
{
    char c1, c2;

    if (case_sensitivity != case_insensitive) return strcmp(s1,s2) != 0;

    for (;;) {
        c1 = *s1;
        c2 = *s2;
        if (c1 >= 65 && c1 <= 90) c1 += 32;
        if (c2 >= 65 && c2 <= 90) c2 += 32;
        if (c1 != c2) return 1;
        if (c1 == 0) return 0;
        ++s1, ++s2;
    }
}

/* LANGUAGE-SPECIFIC IMPROVEMENTS TO DUDE BASED ON CODE REORDERING */

int isHANGUL(u_code_point s) {
    int SIndex = s - 0xAC00;
    if (SIndex < 0 || SIndex >= 11172) {
        return 0;
    }
    return 1;
};
int isUNIHAN(u_code_point s) {
    if (s >= 0x4E00 && s <= 0x9FAF) {
        return 1;
    }
    return 0;
};
int isKATAKANA(u_code_point s) {
    if (s >= 0x30A0 && s <= 0x30FF) {
        return 1;
    }
    return 0;
};
int isHINDI(u_code_point s) {
    if (s >= 0x0900 && s <= 0x0970) {
        return 1;
    }
    return 0;
};
int isLatins(u_code_point s) {
    if (s < 0x370) {
        return 1;
    }
    return 0;
};

// Most frequent 888 Hangeul syllables in Korean BizName
#define HG 888

u_code_point hangeul_freq[HG] = {
    0xd55c, 0xad6d, 0xd559, 0xad50, 0xac00, 0xb8f1, 0xc0c8, 0xc0d0, 0xc0b0, 0xc5c5, 0xacc4, 0xbb3c, 0xb958, 0xc6b4, 0xb3d9, 0xacc2, 0xc775, 0xb8d4, 0xadd8, 0xb2f4, 0xcfe8, 0xc5a8, 0xacc2, 0x0b9a, 0xc811, 0xc2a8, 0xc26c, 0xc9b9d,
};
# define HANGUL_REORDER_BASE 0XB000

u_code_point reorder_hangul(u_code_point s) {
    u_code_point i=HANGUL_REORDER_BASE;
    int k=0;
}
for(k=0; k<HG; k++,i++) {
    if(s == hangeul_freq[k]) { return i; };
};
k=(s - HANGUL_REORDER_BASE);
if( k>=0 && k<HG) {
    return hangeul_freq[k];
};
return s;
}

u_code_point restore_order_hangul(u_code_point z) {
    u_code_point i=HANGUL_REORDER_BASE;
    int k;
    k=(z - HANGUL_REORDER_BASE);
    if( k>=0 && k<HG) {
        return hangeul_freq[k];
    };
    for(k=0; k<HG; k++,i++) {
        if(z == hangeul_freq[k]) { return i; };
    };
    return z;
}

#define UH 4096
u_code_point unihan_freq[UH] = {
    0x4f01,0x696d,0x4e1a,0x5de5,0x7a0b,0x96c6,0x5718,0x56e2,
    0x6709,0x9650,0x8acc,0x88d3,0x4efb,0x516c,0x53f8,0x603b,
    0x90e8,0x767c,0x55d1,0x5c55,0x7ad9,0x70b9,0x958b,0x5f00,
    0x79d1,0x6280,0x8853,0x672f,0x54a8,0x8be2,0x5be6,0x5b9e,
    0x901a,0x4fe1,0x606f,0x7cfa,0x7edf,0x7d71,0x7db2,0x8def,
    0x7edc,0x4e2d,0x5fc3,0x7f51,0x56fd,0x570b,0x9650,0x83ef,
    0x96fb,0x7535,0x5b50,0x8111,0x8166,0x6c23,0x6c14,0x566b,
    0x6a5f,0x6c17,0x529b,0x673a,0x68b0,0x8ba4,0x84d4,0x8a71,
    0x88be,0x8ba1,0x8a2d,0x8a08,0x5099,0x5408,0x52d5,0x52a8,
    0x5236,0x88fd,0x9020,0x4f5c,0x5907,0x8f66,0x8eca,0x975e,
    0x97df,0x8d41,0x5ee0,0x79ae,0x5168,0x7403,0x5c08,0x7523,
    0x773c,0x955c,0x7f8e,0x5bb9,0x6c7d,0x8f66,0x8eca,0x975e,
    0x4f04,0x6362,0x5fcb,0x5916,0x85cf,0x4e91,0x9655,0x8a9e,
    0x57df,0x540d,0x6ce8,0x518c,0x5e7f,0x64ad,0x5ee3,0x544a,
    0x4e3b,0x69e0,0x50b3,0x4e92,0x806f,0x8054,0x5149,0x6750,
    0x5927,0x5b66,0x5b78,0x5206,0x682a,0x5f0f,0x76df,0x534f,
    0x59d4,0x5458,0x4f1a,0x6703,0x793e,0x7559,0x5354,0x6559,
    0x519c,0x4e13,0x51fa,0x7248,0x6587,0x5316,0x827a,0x85dd,
    0x4f53,0x6210,0x4eab,0x624d,0x65e5,0x672c,0x9577,0x6708,
    0x751f,0x6c5d,0x5f8b,0x5e08,0x5e2b,0x8303,0x533b,0x7597,
    0x6cbb,0x836f,0x4fdd,0x5065,0x5e7b,0x8eab,0x967a,0x96aa,
    0x8d38,0x6613,0x8a10,0x4efd,0x8f6f,0x4ef6,0x8edf,0x9ad4,
    0x5a22,0x8cfc,0x7269,0x68d1,0x65c5,0x904a,0x97f3,0x6a02,
    0x670d,0x52d9,0x52a1,0x5546,0x4e8b,0x7814,0x7a76,0x6240,
    0x9662,0x88dc,0x7f6d,0x73ed,0x5100,0x60c5,0x5831,0x9023,
    0x987e,0x95ee,0x514d,0x88d9,0x53f0,0x6e7e,0x8b4a,0x8bc1,
    0x8c0d,0x7e3d,0x5834,0x573a,0x8fb2,0x89c6,0x8288,0x8cbb,
    0x91d1,0x5c5e,0x5c6c,0x92fc,0x6a21,0x9435,0x77be,0x5bc6,
    0x66f8,0x5c4b,0x6175,0x5e97,0x878d,0x904b,0x8f38,0x5ba2,
    0x88b4,0x5238,0x6295,0x8a17,0x9867,0x554f,0x7d9c,0x8ca1,
    0x7d93,0x77ef,0x7968,0x9280,0x884c,0x92b7,0x7b1a,0x7406,
    0x8cb8,0x6b3e,0x5c0f,0x57f8,0x81ea,0x8aea,0x8a62,0x5275,
    0x5bb6,0x5177,0x767e,0x8ca8,0x74b0,0x5883,0x76d1,0x5229,
u_code_point reorder_unihan(u_code_point s) {
    u_code_point i=UNIHAN_REORDER_BASE;
    int k=0;
    for(k=0; k<UH; k++,i++) {
        if(s == unihan_freq[k]) { return i; }
    }
}
#define UNIHAN_REORDER_BASE 0X5000

0x6652,0x8cbf,0x9a74,0x58f3,0x6ca7,0x79a6,0x8d1e,0x818a,
0x8e34,0x6bfc,0x860b,0x71e5,0x5b7d,0x64bc,0x4fcf,0x8f85,
0x79be,0x7538,0x595a,0x8511,0x55d1,0x8d26,0x76ea,0x5be5,
0x66c6,0x8403,0x532f,0x8a98,0x5366,0x557c,0x6361,0x560ed,
0x599e,0x5636,0x555e,0x8b19,0x7caa,0x9470,0x9215,0x6ee5,
0x5b7d,0x64bc,0x4fcf,0x8f85,0x79be,0x7538,0x595a,0x8511,0x55d1,0x8d26,0x76ea,0x5be5,
0x66c6,0x8403,0x532f,0x8a98,0x5366,0x557c,0x6361,0x560ed,
0x599e,0x5636,0x555e,0x8b19,0x7caa,0x9470,0x9215,0x6ee5,
0x5b7d,0x64bc,0x4fcf,0x8f85,0x79be,0x7538,0x595a,0x8511,0x55d1,0x8d26,0x76ea,0x5be5,
0x66c6,0x8403,0x532f,0x8a98,0x5366,0x557c,0x6361,0x560ed,
0x599e,0x5636,0x555e,0x8b19,0x7caa,0x9470,0x9215,0x6ee5,
k=(s - UNIHAN_REORDER_BASE);
if( k>=0 && k<UH) {
    return unihan_freq[k];
}
return s;
}

u_code_point restore_order_unihan(u_code_point z) {
    u_code_point i=UNIHAN_REORDER_BASE;
    int k;
    k=(z - UNIHAN_REORDER_BASE);
    if( k>=0 && k<UH) {
        return unihan_freq[k];
    }
    for(k=0; k<UH; k++,i++) {
        if(z == unihan_freq[k]) { return i; };}
    return z;
}

#define KATAKANA_REORDER_BASE 0X30A0
//KATAKANA reorder by frequency in Japanese Business
#define KK 96
u_code_point katakana_freq[KK] = {
    0x30f3,0x30eb,0x30b9,0x30c8,0x30a2,0x30a4,0x30e9,0x30ea,
    0x30af,0x30c3,0x30fc,0x30b7,0x30b8,0x30e7,0x30ec,0x30b0,
    0x30d5,0x30d7,0x30df,0x30c4,0x30ef,0x30a8,0x30cb,0x30e1,
    0x30ab,0x30c6,0x30b3,0x30dd,0x30d9,0x30cf,0x30c9,0x30a6,
    0x30bb,0x30ce,0x30ca,0x30e0,0x30ed,0x30bf,0x30c1,0x30d0,
    0x30b4,0x30dc,0x30bd,0x30d4,0x30d2,0x30d8,0x30da,0x30db,0x30e3,
    0x30e4,0x30e6,0x30ee,0x30f0,0x30f1,0x30f2,0x30f4,0x30f5,
    0x30f6,0x30f7,0x30f8,0x30f9,0x30fa,0x30fd,0x30ff,0x30a0,
};

u_code_point reorder_katakana(u_code_point s) {
    u_code_point i=KATAKANA_REORDER_BASE;
    int k=0;
    for(k=0; k<KK; k++,i++) {
        if(s == katakana_freq[k]) { return i; };}
    return s; // not reached here
}

u_code_point restore_order_katakana(u_code_point z) {
    return katakana_freq[z - KATAKANA_REORDER_BASE];
}

#define HINDI_REORDER_BASE 0X0900
//HINDI reorder by frequency
#define HD 113
u_code_point hindi_freq[HD] = {
    0x0902,0x093f,0x0940,0x0947,0x0948,0x094b,0x094d,0x0915,
    0x0917,0x0932,0x0938,0x0939,0x0924,0x0926,0x0928,0x092c,
    0x092f,0x0900,0x0901,0x0903,0x0904,0x0916,0x0918,0x0919,
    0x091a,0x091b,0x091c,0x091d,0x091e,0x091f,0x0920,0x0921,
u_code_point reorder_hindi(u_code_point s) {
    u_code_point i=HINDI_REORDER_BASE;
    int k=0;
    for(k=0; k<HD; k++,i++) {
        if(s == hindi_freq[k]) { return i; };
    }
    return s; // not reached here
}

u_code_point restore_order_hindi(u_code_point z) {
    return hindi_freq[z - HINDI_REORDER_BASE];
}

#define MAPCHAR(x,A,B,bytes) if(A<=x && x< (A+bytes)) \
    return(x+(B-A)); if(B<=x && x< (B+bytes)) return(x+(A-B))
#define MAP16BL(x,A,B,block) if(A<=x && x< (A+(block<<4))) \
    return(x+(B-A)); if(B<=x && x< (B+(block<<4))) \
    return(x+(A-B))

u_code_point reorder_latins(u_code_point s) {
    MAP16BL(s,0x0100,0x0000,3); // Latin Extension A
    MAP16BL(s,0x0130,0x0080,2);
    MAP16BL(s,0x0150,0x00A0,1);
    MAP16BL(s,0x0300,0x00B0,3); // Combining Diacritical Marks
    MAPCHAR(s,0x0070,0x0060,1); // p,’
    MAPCHAR(s,0x0072,0x006A,1); // r,j
    MAPCHAR(s,0x0073,0x006B,1); // s,k
    MAPCHAR(s,0x0074,0x0066,1); // t,f
    MAPCHAR(s,0x0075,0x0067,1); // u,g
    MAPCHAR(s,0x0050,0x0040,1); // P,Ø UPPER
    MAPCHAR(s,0x0052,0x004A,1); // R,J UPPER
    MAPCHAR(s,0x0053,0x004B,1); // S,K UPPER
    MAPCHAR(s,0x0054,0x0046,1); // T,F UPPER
    MAPCHAR(s,0x0055,0x0047,1); // U,G UPPER
    MAPCHAR(s,0x0160,0x003A,6); // Latin Extension A
    MAPCHAR(s,0x0166,0x005B,5);
    MAPCHAR(s,0x016B,0x007B,5);
    return s;
}

u_code_point restore_order_latins(u_code_point z) {
    return reorder_latins(z);
}

u_code_point reorder(u_code_point s) {
    if(isHANGUL(s)) return reorder_hangul(s);
    if(isUNIHAN(s)) return reorder_unihan(s);
    return s; // not reached here
}
if (isKATAKANA(s)) return reorder_katakana(s);
if (isHINDI(s)) return reorder_hindi(s);
if (isLatins(s)) return reorder_latins(s);
return s;
}

u_code_point restore_order(u_code_point s) {
    if (isHANGUL(s)) return restore_order_hangul(s);
    if (isUNIHAN(s)) return restore_order_unihan(s);
    if (isKATAKANA(s)) return restore_order_katakana(s);
    if (isHINDI(s)) return restore_order_hindi(s);
    if (isLatins(s)) return restore_order_latins(s);
    return s;
}

/* Encoder: */

enum dude_status dude_encode(
    unsigned int input_length,
    const u_code_point input[],
    const unsigned char uppercase_flags[],
    unsigned int *output_size,
    char output[] )
{
    unsigned int max_out, in, out, k, j;
    u_code_point prev, codept, diff, tmp;
    char shift;

    prev = 0x60;
    //prev = 0x5000;
    max_out = *output_size;

    for (in = out = 0;  in < input_length;  ++in) {
        /* At the start of each iteration, in and out are the number of */
        /* items already input/output, or equivalently, the indices of  */
        /* the next items to be input/output.                           */

        codept = input[in];

        if (codept == 0x2D) {
            /* Hyphen-minus stands for itself. */
            if (max_out - out < 1) return dude_big_output;
            output[out++] = 0x2D;
            continue;
        }

        codept = reorder(codept); // by LSB

        diff = prev^codept;

        /* Compute the number of base-32 characters (k): */
        for (tmp = diff >> 4, k = 1;  tmp != 0;  ++k, tmp >>= 4);
        fprintf(stderr,"diff %x,%x = prev %x ^ codept %x \n",
            k, diff, prev, codept);

        if (max_out - out < k) return dude_big_output;
        shift = uppercase_flags && uppercase_flags[in] ? 32 : 0;
        /* shift controls the case of the last base-32 digit. */

        /*...*/
out += k;
output[out - 1] = base32[diff & 0xF] - shift;
for (j = 2;  j <= k;  ++j) {
    diff >>= 4;
    output[out - j] = base32[0x10 | (diff & 0xF)];
}
prev = codept;
}
/* Append the null terminator: */
if (max_out - out < 1) return dude_big_output;
output[out++] = 0;
*output_size = out;
return dude_success;
}
/* Decoder: */
enum dude_status dude_decode(
    enum case_sensitivity case_sensitivity,
    char scratch_space[],
    const char input[],
    unsigned int *output_length,
    u_code_point output[],
    unsigned char uppercase_flags[]
){
    u_code_point prev, q, diff;
    char c;
    unsigned int max_out, in, out, scratch_size;
    enum dude_status status;
    prev = 0x60;
    max_out = *output_length;
    for (c = input[in = 0], out = 0;  c != 0;  c = input[++in], ++out) {
        /* At the start of each iteration, in and out are the number of */
        /* items already input/output, or equivalently, the indices of  */
        /* the next items to be input/output.                           */
        if (max_out - out < 1) return dude_big_output;

        if (c == 0x2D) output[out] = c;  /* hyphen-minus is literal */
        else {
            /* Base-32 sequence. Decode quintets until 0xxxx is found: */

            for (diff = 0;  ;  c = input[++in]) {
                q = base32_decode(c);
                if (q == base32_invalid){ return dude_bad_input; }
                diff = (diff << 4) | (q & 0xF);
                if (q >> 4 == 0) break;
            }
            // prev = output[out] = prev ^ diff;
prev = prev ^ diff;
output[out] = restore_order(prev); // LSB
}

/* Case of last character determines uppercase flag: */
if (uppercase_flags) uppercase_flags[out] = c >= 65 && c <= 90;

/* Enforce the uniqueness of the encoding by re-encoding */
/* the output and comparing the result to the input: */
scratch_size = ++in;
status = dude_encode(out, output, uppercase_flags,
                    &scratch_size, scratch_space);
if (status != dude_success || scratch_size != in ||
    unequal(case_sensitivity, scratch_space, input))
    return dude_bad_input;
*output_length = out;
return dude_success;
}

/***************************************************************************/
/* Wrapper for testing (would normally go in a separate .c file): */
#include <assert.h>
#include <stdio.h>
#include <stdlib.h>
#include <string.h>

/* For testing, we'll just set some compile-time limits rather than */
/* use malloc(), and set a compile-time option rather than using a */
/* command-line option. */
enum {
    unicode_max_length = 256,
    ace_max_size = 256,
    test_case_sensitivity = case_insensitive
        /* suitable for host names */
};

static void usage(char **argv)
{
    fprintf(stderr,
            "%s -e reads code points and writes a DUDE string.\n" "%s -d reads a DUDE string and writes code points.\n" "Input and output are plain text in the native character set.\n" "Code points are in the form u+hex separated by whitespace.\n" "A DUDE string is a newline-terminated sequence of LDH characters\n" "(without any signature).\n" "The case of the u in u+hex is the force-to-uppercase flag.\n" , argv[0], argv[0]);
    exit(EXIT_FAILURE);
}

static void fail(const char *msg)
{
    fputs(msg, stderr);
}
int main(int argc, char **argv) {
    enum dude_status status;
    int r;
    char *p;

    if (argc != 2) usage(argv);
    if (argv[1][0] != '-') usage(argv);
    if (argv[1][2] != 0) usage(argv);

    if (argv[1][1] == 'e') {
        u_code_point input[unicode_max_length];
        unsigned long codept;
        unsigned char uppercase_flags[unicode_max_length];
        char output[ace_max_size], uplus[3];
        unsigned int input_length, output_size, i;

        /* Read the input code points: */
        input_length = 0;

        for (;;) {
            r = scanf("%2s%lx", uplus, &codept);
            if (ferror(stdin)) fail(io_error);
            if (r == EOF || r == 0) break;

            if (r != 2 || uplus[1] != '+' || codept > (u_code_point)-1) {
                fail(invalid_input);
            }

            if (input_length == unicode_max_length) fail(too_big);

            if (uplus[0] == 'u') uppercase_flags[input_length] = 0;
            else if (uplus[0] == 'U') uppercase_flags[input_length] = 1;
            else fail(invalid_input);

            input[input_length++] = codept;
        }
    }
}
/* Encode: */

output_size = ace_max_size;
status = dude_encode(input_length, input, uppercase_flags,
       &output_size, output);
if (status == dude_bad_input) fail(invalid_input);
if (status == dude_big_output) fail(too_big);
assert(status == dude_success);

/* Convert to native charset and output: */

for (p = output;  *p != 0;  ++p) {
   i = *p;
   assert(i <= 122 && ldh_ascii[i] != '.');
   *p = ldh_ascii[i];
}

r = puts(output);
fprintf(stderr,"length: %d\n", strlen(output));
if (r == EOF) fail(io_error);
return EXIT_SUCCESS;

if (argv[1][1] == 'd') {
   char input[ace_max_size], scratch[ace_max_size], *pp;
   u_code_point output[unicode_max_length];
   unsigned char uppercase_flags[unicode_max_length];
   unsigned int input_length, output_length, i;

   /* Read the DUDE input string and convert to ASCII: */

   fgets(input, ace_max_size, stdin);
   if (ferror(stdin)) fail(io_error);
   if (feof(stdin)) fail(invalid_input);
   input_length = strlen(input);
   if (input[input_length - 1] != '\n') fail(too_big);
   input[--input_length] = 0;

   for (p = input;  *p != 0;  ++p) {
      pp = strchr(ldh_ascii, *p);
      if (pp == 0) fail(invalid_input);
      *p = pp - ldh_ascii;
   }

   /* Decode: */

   output_length = unicode_max_length;
   status = dude_decode(test_case_sensitivity, scratch, input,
                   &output_length, output, uppercase_flags);
   if (status == dude_bad_input) fail(invalid_input);
   if (status == dude_big_output) fail(too_big);
   assert(status == dude_success);

   /* Output the result: */

   for (i = 0;  i < output_length;  ++i) {
      r = printf(}%s+%04lX\n",
                     uppercase_flags[i] ? "U" : "u",
       (unsigned long) output[i] );
      if (r < 0) fail(io_error);
return EXIT_SUCCESS;
}

usage(argv);
return EXIT_SUCCESS; /* not reached, but quiets compiler warning */
}

LAMCW: Example implementation into AMC-ACE-W

This idea is applicable to any ACEs.
LAMCW is a name for AMC-ACE-W implementation of this idea.

Embedded hangul, han and Latin frequency tables are subject to change with further studies in the next revision of this draft.

In Unix, save this example source code into ldude.c

% cc -o lamcw lamcw.c
% ./lamcw -e < input_file > output_file
% ./lamcw -d < output_file

An input file should contains u+????-form code points delimited with spaces or newlines.

/*********************/
/* lamcw.c 1.0 (2001-Jul-3) */
/* Soobok Lee <lsb@postel.co.kr> */
/* amcw.c from Adam M. Costello <amc@cs.berkeley.edu> */
/*********************/

/*********************/
/* This is ANSI C code (C89) implementing AMC-ACE-W version 0.1.*. */
/*********************/
/* Public interface (would normally go in its own .h file): */

#include <limits.h>

enum amc_ace_status {
    amc_ace_success,
    amc_ace_bad_input,
    amc_ace_big_output /* Output would exceed the space provided. */
};

enum case_sensitivity { case_sensitive, case_insensitive };

#if UINT_MAX >= 0x1FFFFF
typedef unsigned int u_code_point;
#else
typedef unsigned long u_code_point;
#endif

enum amc_ace_status amc_ace_w_encode(
    unsigned int input_length,
    const u_code_point input[],
    const unsigned char uppercase_flags[],
    unsigned int *output_size,
    char output[]);

/* amc_ace_w_encode() converts Unicode to AMC-ACE-W (without any signature). The input must be represented as an array of Unicode code points (not code units; surrogate pairs are not allowed), and the output will be represented as null-terminated ASCII. The input_length is the number of code points in the input. The output_size is an in/out argument: the caller must pass in the maximum number of characters that may be output (including the terminating null), and on successful return it will contain the number of characters actually output (including the terminating null, so it will be one more than strlen() would return, which is why it is called output_size rather than output_length). The uppercase_flags array must hold input_length boolean values, where nonzero means the corresponding Unicode character should be forced to uppercase after being decoded, and zero means it is caseless or should be forced to lowercase. Alternatively, uppercase_flags may be a null pointer, which is equivalent to all zeros. The letters a-z and A-Z are always encoded literally, regardless of the corresponding flags. The encoder always outputs lowercase base-32 characters except when nonzero values of uppercase_flags require otherwise. The return value may be any of the amc_ace_status values defined above; if not amc_ace_success, then output_size and output may contain garbage. On success, the encoder will never need to write an output_size greater than input_length*5+1, because of how the encoding is defined.

enum amc_ace_status amc_ace_w_decode(
    enum case_sensitivity case_sensitivity,
    char scratch_space[],
    const char input[],
    unsigned int *output_length,
    u_code_point output[],
    unsigned char uppercase_flags[]);

/* amc_ace_w_decode() converts AMC-ACE-W (without any signature) to Unicode. The input must be represented as null-terminated ASCII, and the output will be represented as an array of Unicode code points. The case_sensitivity argument influences the check on the well-formedness of the input string; it must be case_sensitive if case-sensitive comparisons are allowed on encoded strings, case_insensitive otherwise. The scratch_space must point to space at least as large as the input, which will get overwritten (this allows the decoder to avoid calling malloc()). The output_length is an in/out argument: the caller must pass in the maximum number of code points that may be output, and on successful return it will contain the actual number of code points output. The uppercase_flags array must have room for at least output_length values, or it may be a null pointer if the case information is not needed. A nonzero flag indicates that the corresponding Unicode character should be forced to uppercase by the caller, while zero means it is caseless or should be forced to lowercase. The letters a-z and A-Z are always encoded literally, regardless of the corresponding flags. The decoder always outputs lowercase base-32 characters except when nonzero values of uppercase_flags require otherwise. The return value may be any of the amc_ace_status values defined above; if not amc_ace_success, then output_size and output may contain garbage. On success, the encoder will never need to write an output_size greater than input_length*5+1, because of how the encoding is defined. */
/* a-z and A-Z are output already in the proper case, but their flags will be set appropriately so that applying the flags would be harmless. The return value may be any of the amc_ace_status values defined above; if not amc_ace_success, then output_length, output, and uppercase_flags may contain garbage. On success, the decoder will never need to write an output_length greater than the length of the input (not counting the null terminator), because of how the encoding is defined. */

/**********************************************************/
/* Implementation (would normally go in its own .c file): */
#
#include <string.h>

/* base32[q] is the lowercase base-32 character representing the number q from the range 0 to 31. Note that we cannot use string literals for ASCII characters because an ANSI C compiler does not necessarily use ASCII. */

static const char base32[] = {
    97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, /* a-k */
    109, 110, /* m-n */
    112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, /* p-z */
    50, 51, 52, 53, 54, 55, 56, 57 /* 2-9 */
};

/* base32_decode(c) returns the value of a base-32 character, in the range 0 to 31, or the constant base32_invalid if c is not a valid base-32 character. */

enum { base32_invalid = 32 };

static unsigned int base32_decode(char c)
{
    if (c < 50) return base32_invalid;
    if (c <= 57) return c - 26;
    if (c < 97) c += 32;
    if (c < 97 || c == 108 || c == 111 || c > 122) return base32_invalid;
    return c - 97 - (c > 108) - (c > 111);
}

/* unequal(case_sensitivity,s1,s2) returns 0 if the strings s1 and s2 are equal, 1 otherwise. If case_sensitivity is case_insensitive, then ASCII A-Z are considered equal to a-z respectively. */

static int unequal( enum case_sensitivity case_sensitivity,
    const char s1[], const char s2[] )
{
    char c1, c2;

    if (case_sensitivity != case_insensitive) return strcmp(s1,s2) != 0;

    for (;;) {
        c1 = *s1;
        c2 = *s2;
        if (c1 >= 65 && c1 <= 90) c1 += 32;
        if (c2 >= 65 && c2 <= 90) c2 += 32;
        if (c1 != c2) return 1;
        if (c1 == 0) return 0;
    }
++s1, ++s2;
}
}

/* LANGUAGE-SPECIFIC IMPROVEMENTS TO DUDE BASED ON CODE REORDERING */

int isHANGUL(u_code_point s) {
    int SIndex = s - 0xAC00;
    if (SIndex < 0 || SIndex >= 11172) {
        return 0;
    }
    return 1;
};

int isUNIHAN(u_code_point s) {
    if (s >= 0x4E00 && s <= 0x9FAF) {
        return 1;
    }
    return 0;
};

int isKATAKANA(u_code_point s) {
    if (s >= 0x30A0 && s <= 0x30FF) {
        return 1;
    }
    return 0;
};

int isHINDI(u_code_point s) {
    if (s >= 0x0900 && s <= 0x0970) {
        return 1;
    }
    return 0;
};

int isLatins(u_code_point s) {
    if (s < 0x370) {
        return 1;
    }
    return 0;
};

#define HG 888

u_code_point hangeul_freq[HG] = {
    0xd55c, 0xad6d, 0xd559, 0xad50, 0xb300, 0xce00, 0xb4f1, 0xc0d8,
    0xc911, 0xb824, 0xd654, 0xd604, 0xc6d0, 0xbb38, 0xc721, 0xbe1d,
    0xc804, 0xc790, 0xae30, 0xc5f5, 0xc0b0, 0xc5c5, 0xacc4, 0xbb3c,
    0xb958, 0xc6b4, 0xb3d9, 0xcc28, 0xc220, 0xd56d, 0xb3e8, 0xc68d,
    0xc74, 0xc1d4, 0xae8e, 0xd305, 0xacc5, 0xc0d8, 0xc774, 0xc368,
    0xc11c, 0xc6b8, 0xbe44, 0xc2dc, 0xc1a4, 0xfd15, 0xd14d, 0xda0d,
    0xcc8c, 0xd22d, 0xda3e, 0xce20, 0xbc30, 0xda2e, 0xc68b, 0xc2f3,
    0xc815, 0xbc4e, 0xda5b, 0xc2e0, 0xc0c1, 0xc0ac, 0xda6c, 0xc138,
    0xc6a9, 0xda11, 0xda6c, 0xda30, 0xc9c4, 0xb791, 0xb9e4, 0xda5d,
    0xd004, 0xda54, 0xc1f4, 0xda51, 0xb0a0, 0xb110, 0xc370, 0xc774,
    0xda48, 0xb9c8, 0xda14, 0xda6c, 0xc0bf, 0xc9d0, 0xc2ed, 0xc548,
    0xc18c, 0xd504, 0xda2b, 0xc6e8, 0xbfb8, 0xb514, 0xc5b4, 0xc544,
    0xda53, 0xda30, 0xc5f4, 0xb9ac, 0xc7e4, 0xc7c7, 0xc5f0, 0xb5a4,
    0xda22c, 0xc740, 0xda59, 0xc7e4, 0xc4d4, 0xda8f, 0xb8f9, 0xb9b0,
    0xc6d4, 0xb79c, 0xc5ec, 0xc88b, 0xce8e, 0xc9e0, 0xb9bc, 0xda578,
```c
u_code_point reorder_hangul(u_code_point s) {
    int k=0;
    for(k=0; k<HG; k++,i++) {
        if(s == hangeul_freq[k]) { return i; };
    }
    k=(s - HANGUL_REORDER_BASE);
    if( k>=0 && k<HG) {
        return hangeul_freq[k];
    }
    return s;
}

u_code_point restore_order_hangul(u_code_point z) {
    int k=0;
    k=(z - HANGUL_REORDER_BASE);
    if(k>=0 && k<HG) {
        return hangeul_freq[k];
    }
    for(k=0; k<HG; k++,i++) {
        if(z == hangeul_freq[k]) { return i; };
    }
    #define HANGUL_REORDER_BASE 0XB000
```
```c
};
return z;
}

// Most frequent 4096 SC/TC characters in CJK
#define UH 4096
u_code_point unihan_freq[UH] = {
0x4f01, 0x696d, 0x4e1a, 0x5de5, 0x7a0b, 0x96c6, 0x56e2, 0x7609, 0x9650, 0x8cac, 0x8d23, 0x4efb, 0x516c, 0x53f8, 0x603b, 0x90e8, 0x767c, 0x53d1, 0x5c55, 0x7a0b, 0x97d1, 0x6280, 0x8853, 0x5718, 0x56e2, 0x901a, 0xe4f1, 0x606f, 0x77f1, 0x7d71, 0x7db2, 0x8def, 0x7edf, 0x8eab, 0x9650, 0x8111, 0x8166, 0x6c23, 0x6c14, 0x5688, 0x56f1, 0x6c17, 0x529b, 0x673a, 0x8ba1, 0x7fd2, 0x73ed, 0x5100, 0x60c5, 0x83b1, 0x827a, 0x7fd2, 0x73ed, 0x5100, 0x60c5, 0x5831, 0x9023, 0x987e, 0x95ee, 0x514d, 0x8d39, 0x53f0, 0x6e7e, 0x8ba4, 0x8bc1, 0x8d58, 0x52d2, 0x52a1, 0x5546, 0x4e16, 0x6cb3, 0x6c5f, 0x6e56, 0x7701, 0x5317, 0x897f, 0x6b21, 0x5d8b, 0x798f, 0x5ca1, 0x5bae, 0x8a0a, 0x8cc7, 0x7532, 0x753a, 0x6851, 0x6597, 0x6a4b, 0x7532, 0x6d5c, 0x718a, 0x623f, 0x7b2c, 0x79cb, 0x8218, 0x4e0a, 0x962a, 0x52dd,
```
#define UNIHAN_REORDER_BASE 0X5000

u_code_point reorder_unihan(u_code_point s) {
    u_code_point i=UNIHAN_REORDER_BASE;
    int k=0;
    for(k=0; k<UH; k++,i++) {
        if(s == unihan_freq[k]) { return i; };
    };
    k=(s - UNIHAN_REORDER_BASE);
    if( k>=0 && k<UH) {
        return unihan_freq[k];
    };
    return s;
}

u_code_point restore_order_unihan(u_code_point z) {
    u_code_point i=UNIHAN_REORDER_BASE;
    int k;
    k=(z - UNIHAN_REORDER_BASE);
    if( k>=0 && k<UH) {
        return unihan_freq[k];
    };
    for(k=0; k<UH; k++,i++) {
        if(z == unihan_freq[k]) { return i; };
    };
    return z;
}
#define KATAKANA_REORDER_BASE 0x30A0
// KATAKANA reorder by frequency in Japanese Business
#define KK 96
u_code_point katakana_freq[KK] = {
    0x30f3, 0x30eb, 0x30b9, 0x30c8, 0x30a2, 0x3004, 0x30e7, 0x30ec, 0x30b0, 0x30d5, 0x307d, 0x30df, 0x30c4, 0x30ef, 0x30a8, 0x30cb, 0x30e1, 0x30ab, 0x30c6, 0x30b3, 0x30dd, 0x30d9, 0x30cf, 0x30c9, 0x30a6, 0x30bb, 0x30ce, 0x30ca, 0x30e0, 0x30ed, 0x30bf, 0x30c1, 0x30d0, 0x30b4, 0x30dc, 0x30bd, 0x30cd, 0x30e2, 0x30d3, 0x30b5, 0x30ad, 0x30b1, 0x30b2, 0x30bc, 0x30e8, 0x30e5, 0x30aa, 0x30cc, 0x30a3, 0x30cd, 0x30de, 0x30a1, 0x30a5, 0x30a7, 0x30a9, 0x30ac, 0x30ae, 0x30b6, 0x30ba, 0x30be, 0x30c0, 0x30c2, 0x30c5, 0x30c7, 0x30d1, 0x30d2, 0x30d4, 0x30d8, 0x30da, 0x30db, 0x30e3, 0x30e4, 0x30e6, 0x30ee, 0x30f0, 0x30f1, 0x30f2, 0x30f4, 0x30f5, 0x30f6, 0x30f7, 0x30f8, 0x30f9, 0x30fa, 0x30fb, 0x30fd, 0x30fe, 0x30ff, 0x30a0,
};

u_code_point reorder_katakana(u_code_point s) {
    u_code_point i = KATAKANA_REORDER_BASE;
    int k = 0;
    for (k = 0; k < KK; k++, i++) {
        if (s == katakana_freq[k]) { return i; }
    }
    return s; // not reached here
}

u_code_point restore_order_katakana(u_code_point z) {
    return katakana_freq[z - KATAKANA_REORDER_BASE];
}

#define HINDI_REORDER_BASE 0x0900
// HINDI reorder by frequency in Japanese Business
#define HD 113
u_code_point hindi_freq[HD] = {
    0x0902, 0x093f, 0x0940, 0x0947, 0x0948, 0x094b, 0x094d, 0x0915, 0x0917, 0x0932, 0x0938, 0x0939, 0x0924, 0x0926, 0x0928, 0x092c, 0x092f, 0x0900, 0x0901, 0x0903, 0x0904, 0x0916, 0x0918, 0x0919, 0x091a, 0x091b, 0x091c, 0x091d, 0x091e, 0x091f, 0x0920, 0x0921, 0x0922, 0x0923, 0x0925, 0x0927, 0x0929, 0x092a, 0x092b, 0x092d, 0x092e, 0x0930, 0x0931, 0x0933, 0x0934, 0x0935, 0x0936, 0x0937, 0x093a, 0x093b, 0x093c, 0x093d, 0x093e, 0x0941, 0x0942, 0x0943, 0x0944, 0x0945, 0x0946, 0x0949, 0x094a, 0x094c, 0x094e, 0x094f, 0x0950, 0x0951, 0x0952, 0x0953, 0x0954, 0x0955, 0x0956, 0x0957, 0x0958, 0x0959, 0x095a, 0x095b, 0x095c, 0x095d, 0x095e, 0x095f, 0x0960, 0x0961, 0x0962, 0x0963, 0x0964, 0x0965, 0x0966, 0x0967, 0x0968, 0x0969, 0x096a, 0x096b, 0x096c, 0x096d, 0x096e, 0x096f, 0x0905, 0x0906, 0x0907, 0x0908, 0x0909, 0x090a, 0x090b, 0x090c, 0x090d, 0x090e, 0x090f, 0x0910, 0x0911, 0x0912, 0x0913, 0x0914, 0x0970,
};

u_code_point reorder_hindi(u_code_point s) {
    u_code_point i = HINDI_REORDER_BASE;
    int k = 0;
    for (k = 0; k < HD; k++, i++) {
        if (s == hindi_freq[k]) { return i; }
    }
    return s; // not reached here
u_code_point restore_order_hindi(u_code_point z) {
    return hindi_freq[z - HINDI_REORDER_BASE];
}

#define MAPCHAR(x,A,B,bytes) if(A<=x && x< (A+bytes)) return(x+(B-A)); if(B<=x && x< (B+bytes)) return(x+(A-B))

#define MAP16BL(x,A,B,block) if(A<=x && x< (A+(block<<4))) return(x+(B-A)); if(B<=x && x< (B+(block<<4))) return(x+(A-B))

u_code_point reorder_latins(u_code_point s) {
    MAP16BL(s,0x0100,0x0000,3); // Latin Extension A
    MAP16BL(s,0x0130,0x0080,2);
    MAP16BL(s,0x0150,0x00A0,1);
    MAP16BL(s,0x0300,0x00B0,3); // Combining Diacritical Marks
    MAPCHAR(s,0x0070,0x0060,1); // p,'
    MAPCHAR(s,0x0072,0x006A,1); // r,j
    MAPCHAR(s,0x0073,0x006B,1); // s,k
    MAPCHAR(s,0x0074,0x0066,1); // t,f
    MAPCHAR(s,0x0075,0x0067,1); // u,g
    MAPCHAR(s,0x0050,0x0040,1); // P,@ UPPER
    MAPCHAR(s,0x0052,0x004A,1); // R,J UPPER
    MAPCHAR(s,0x0053,0x004B,1); // S,K UPPER
    MAPCHAR(s,0x0054,0x0046,1); // T,F UPPER
    MAPCHAR(s,0x0055,0x0047,1); // U,G UPPER
    MAPCHAR(s,0x0160,0x003A,6); // Latin Extension A
    MAPCHAR(s,0x0166,0x005B,5);
    MAPCHAR(s,0x016B,0x007B,5);
    return s;
}

u_code_point restore_order_latins(u_code_point z) { return reorder_latins(z); }

u_code_point reorder(u_code_point s) {
    if(isHANGUL(s)) return reorder_hangul(s);
    if(isUNIHAN(s)) return reorder_unihan(s);
    if(isKATAKANA(s)) return reorder_katakana(s);
    if(isHINDI(s)) return reorder_hindi(s);
    if(isLatins(s)) return reorder_latins(s);
    return s;
}

u_code_point restore_order(u_code_point s) {
    if(isHANGUL(s)) return restore_order_hangul(s);
    if(isUNIHAN(s)) return restore_order_unihan(s);
    if(isKATAKANA(s)) return restore_order_katakana(s);
    if(isHINDI(s)) return restore_order_hindi(s);
    if(isLatins(s)) return restore_order_latins(s);
    return s;
}

/* update(refpoint,style,n,k) updates refpoint[1..3] and *style */
/** based on n (the most recent code point) and k (the number of */
/* base-32 characters used to encode it). */

static void update( u_code_point refpoint[6], unsigned int *style,
        u_code_point n, unsigned int k )
{
    *style = k < 3 ? 0 : k > 3 ? 1 : *style;
    refpoint[1] = (n >> 4) << 4;
    if (k > 2) refpoint[2] = n - 0xA0 < 0xE0 ? 0xA0 : (n >> 8) << 8;
    if (k > 3) refpoint[3] = n - 0x3000 < 0x7000 ? 0x4E00 :
        *style == 1 && n - 0xA000 < 0x3800 ? 0x8800 : (n >> 12) << 12;
}

/* Main encode function: */

enum amc_ace_status amc_ace_w_encode(
    unsigned int input_length,
    const u_code_point input[],
    const unsigned char uppercase_flags[],
    unsigned int *output_size,
    char output[] )
{
    unsigned int style, literal, max_out, in, out, k, j;
    u_code_point n, delta;
    const u_code_point maxdelta[2][6] =
        {{0,0xF,0xFF,0xFFFF,0xFFFFF,0xFFFFF}, {0,0,0xFF,0x4FFF,0xFFFF,0xFFFFF}};
    char shift;

    /* Initialize the state: */
    u_code_point refpoint[6] = {0, 0xE0, 0xA0, 0, 0, 0x10000};
    style = literal = 0;
    max_out = *output_size;

    for (in = out = 0; in < input_length; ++in) {

        /* At the start of each iteration, in and out are the number of */
        /* items already input/output, or equivalently, the indices of */
        /* the next items to be input/output. */

        n = input[in];
        /* Check the code point range to avoid array bounds errors later: */
        if (n > 0x10FFFF) return amc_ace_bad_input;

        if (n == 0x2D) {
            /* Hyphen-minus is doubled. */
            if (max_out - out < 2) return amc_ace_big_output;
            output[out++] = 0x2D;
            output[out++] = 0x2D;
        }
        else if (n >= 122 &&
        (n >= 97 ||
        (n >= 65 && n <= 90) ) ) {
            /* Encode an LDH character literally. */
            if (max_out - out < 1 + !literal) return amc_ace_big_output;
            /* Switch to literal mode if necessary: */
            if (!literal) output[out++] = 0x2D;
            literal = 1;
            output[out++] = n;
        }
        else {
/* Encode a non-LDH character using base-32. */
/* First compute the number of base-32 characters (k): */

n = reorder(n); // ADDED *****************

for (k = 1 + style;  ;  ++k) {
    delta = n - refpoint[k];
    if (delta <= maxdelta[style][k]) break;
}

if (max_out - out < k + literal) return amc_ace_big_output;
/* Switch to base-32 mode if necessary: */
if (literal) output[out++] = 0x2D;
literal = 0;
shift = uppercase_flags && uppercase_flags[in] ? 32 : 0;

/* Check for the extended delta of style 1 window 3: */
if (k == 3 && delta >= 0x1000) {
    /* The top 16k of window 3 is encoded as 0xxxx xxxx 000xx. */
    delta -= 0x1000;
    output[out++] = base32[delta >> 10] - shift;
    output[out++] = base32[(delta >> 5) & 0x1F];
    output[out++] = base32[delta & 0x1F];
} else {
    /* Each quintet has the form 1xxxx except the last is 0xxxx. */
    /* Computing the base-32 digits in reverse order is easiest. */
    out += k;
    output[out - 1] = base32[delta & 0xF] - shift;

    for (j = 2;  j <= k;  ++j) {
        delta >>= 4;
        output[out - j] = base32[0x10 | (delta & 0xF)];
    }
}

update(refpoint, &style, n, k);

/* Append the null terminator: */
if (max_out - out < 1) return amc_ace_big_output;
output[out++] = 0;

*output_size = out;
return amc_ace_success;
}

/* Main decode function: */

enum amc_ace_status amc_ace_w_decode(
    enum case_sensitivity case_sensitivity,
    char scratch_space[],
    const char input[],
    unsigned int *output_length,
    u_code_point output[],
    unsigned char uppercase_flags[]
) {
    u_code_point q, delta;
char c;
unsigned int style, literal, max_out, in, out, k, scratch_size;
enum amc_ace_status status;

/* Initialize the state: */

u_code_point refpoint[6] = {0, 0xE0, 0xA0, 0, 0, 0x10000};

style = literal = 0;
max_out = *output_length;

for (c = input[in = 0], out = 0; c != 0; c = input[++in], ++out) {

    /* At the start of each iteration, in and out are the number of */
    /* items already input/output, or equivalently, the indices of   */
    /* the next items to be input/output. c is the same as input[in] */
    /* except when "extra" characters have been consumed (see below). */

    if (c == 0x2D && input[in + 1] != 0x2D) {
        /* Unpaired hyphen-minus toggles mode. */
        literal = !literal;
        c = input[++in];
    }

    if (max_out - out < 1) return amc_ace_big_output;

    if (c == 0x2D) {
        /* Double hyphen-minus represents a hyphen-minus. */
        ++in;
        output[out] = 0x2D;
    } else {
        if (literal) output[out] = c;
        else {
            /* Decode a base-32 sequence.                  */
            /* First decode quintets until 0xxxx is found: */

            for (delta = 0, k = 1; ; c = input[++in], ++k) {
                q = base32_decode(c);
                if (q == base32_invalid || k > 5) return amc_ace_bad_input;
                delta = (delta << 4) | (q & 0xF);
                if (q >> 4 == 0) break;
            }

            if (style == 1 && k == 1) {
                /* Style 1 has no window 1, so it must be the extended */
                /* delta of window 3, encoded as 0xxxx xxxx x.xxx.     */
                /* Consume the two "extra" characters:              */

                for (; k < 3; ++k) {
                    q = base32_decode(input[++in]);
                    if (q == base32_invalid) return amc_ace_bad_input;
                    delta = (delta << 5) | q;
                }

                delta += 0x1000;
            }

            output[out] = refpoint[k] + delta;
            update(refpoint, &style, output[out], k);
        }
    }
}
output[out] = restore_order(output[out]); // ADDED

/* Case of last non-extra character determines uppercase flag: */
if (uppercase_flags) uppercase_flags[out] = c >= 65 && c <= 90;

/* Enforce the uniqueness of the encoding by re-encoding */
/* the output and comparing the result to the input: */
scratch_size = ++in;
status = amc_ace_w_encode(out, output, uppercase_flags,
    &scratch_size, scratch_space);
if (status != amc_ace_success || scratch_size != in ||
    unequal(case_sensitivity, scratch_space, input))
    return amc_ace_bad_input;

*output_length = out;
return amc_ace_success;
}

/***************************************************************************/
/* Wrapper for testing (would normally go in a separate .c file): */

#include <assert.h>
#include <stdio.h>
#include <stdlib.h>
#include <string.h>

/* For testing, we'll just set some compile-time limits rather than */
/* use malloc(), and set a compile-time option rather than using a */
/* command-line option. */
enum {
    unicode_max_length = 256,
    ace_max_size = 256,
    test_case_sensitivity = case_insensitive
        /* suitable for host names */
};

static void usage(char **argv)
{
    fprintf(stderr,
        "\$s -e reads code points and writes an AMC-ACE-W string.\n"\n        "\$s -d reads an AMC-ACE-W string and writes code points.\n"\n        "Input and output are plain text in the native character set.\n"\n        "Code points are in the form u+hex separated by whitespace.\n"\n        "An AMC-ACE-W string is a newline-terminated sequence of LDH\n"\n        "characters (without any signature).\n"\n        "The case of the u in u+hex is the force-to-uppercase flag.\n"\n        , argv[0], argv[0]);
    exit(EXIT_FAILURE);
}

static void fail(const char *msg)
{
    fputs(msg, stderr);
}
exit(EXIT_FAILURE);

static const char too_big[] =
  "input or output is too large, recompile with larger limits\n";
static const char invalid_input[] = "invalid input\n";
static const char io_error[] = "I/O error\n";

/* The following string is used to convert LDH characters between ASCII and the native charset: */

static const char ldh_ascii[] =
  ". ............" 
  ". ............" 
  ". ............-.." 
  "0123456789......" 
  ".ABCDEFGHIJKLMNOPQRSTUVWXYZ....." 
  ".abcdefghijklmnopqrstuvwxyz" 

int main(int argc, char **argv)
{
  enum amc_ace_status status;
  int r;
  char *p;

  if (argc != 2) usage(argv);
  if (argv[1][0] != '-') usage(argv);
  if (argv[1][2] != 0) usage(argv);
  if (argv[1][1] == 'e') {
    u_code_point input[unicode_max_length];
    unsigned long codept;
    unsigned char uppercase_flags[unicode_max_length];
    char output[ace_max_size], uplus[3];
    unsigned int input_length, output_size, i;

    /* Read the input code points: */
    input_length = 0;
    for (;;) {
      r = scanf("%2s%lx", uplus, &codept);
      if (ferror(stdin)) fail(io_error);
      if (r == EOF || r == 0) break;
      if (r != 2 || uplus[1] != '+' || codept > (u_code_point)-1) {
        fail(invalid_input);
      }
      if (input_length == unicode_max_length) fail(too_big);
      if (uplus[0] == 'u') uppercase_flags[input_length] = 0;
      else if (uplus[0] == 'U') uppercase_flags[input_length] = 1;
      else fail(invalid_input);
      input[input_length++] = codept;
    }

    ...
output_size = ace_max_size;
status = amc_ace_w_encode(input_length, input, uppercase_flags,
   &output_size, output);
if (status == amc_ace_bad_input) fail(invalid_input);
if (status == amc_ace_big_output) fail(too_big);
assert(status == amc_ace_success);

for (p = output; *p != 0; ++p) {
   i = *p;
   assert(i <= 122 && ldh_ascii[i] != '.');
   *p = ldh_ascii[i];
}

r = puts(output);
if (r == EOF) fail(io_error);
return EXIT_SUCCESS;
}

if (argv[1][1] == 'd') {
   char input[ace_max_size], scratch[ace_max_size], *pp;
   u_code_point output[unicode_max_length];
   unsigned char uppercase_flags[unicode_max_length];
   unsigned int input_length, output_length, i;
   /* Read the AMC-ACE-W input string and convert to ASCII: */
   fgets(input, ace_max_size, stdin);
   if (ferror(stdin)) fail(io_error);
   if (feof(stdin)) fail(invalid_input);
   input_length = strlen(input);
   if (input[input_length - 1] != '\n') fail(too_big);
   input[--input_length] = 0;
   for (p = input; *p != 0; ++p) {
      pp = strchr(ldh_ascii, *p);
      if (pp == 0) fail(invalid_input);
      *p = pp - ldh_ascii;
   }

   output_length = unicode_max_length;
   status = amc_ace_w_decode(test_case_sensitivity, scratch, input,
      &output_length, output, uppercase_flags);
   if (status == amc_ace_bad_input) fail(invalid_input);
   if (status == amc_ace_big_output) fail(too_big);
   assert(status == amc_ace_success);

   /* Output the result: */
   for (i = 0; i < output_length; ++i) {
      r = printf("%s+%04lX\n",
         uppercase_flags[i] ? "U" : "u",
         (unsigned long) output[i]);
      if (r < 0) fail(io_error);
   }
}
return EXIT_SUCCESS;
}

usage(argv);
return EXIT_SUCCESS; /* not reached, but quiets compiler warning */

/* end of lamcw.c */