dnj-corpus

A small corpus of a local newspaper (*-Pamɛbhamɛ*), and medical counsels (chapters) from *While waiting for a medical doctor* translated into Eastern Dan.

Language Description

- **ISO 639-3 language tag**: dnj
- **Language Name**: Dan
- **Language variety demonstrated in this corpus**: Eastern Dan
- **Script**: Can is written in Latin script.

Language Note: Dan is considered by some to be a macro language comprised of a dialect chain of over 40 dialects [3,4]. As recently as 2012 the ISO 639-3 registrar approved a request (2012-083)[5] to split one of these dialects off into its own language (Kla [lda]). Eastern and Western Dan have had their own separate writing traditions for over 40 years. There are significant segmental and suprasegmental differences between Eastern and Western Dan.

Script Note: There may be several orthographies from different dialects which would all qualify as BCP47[6]: dnj-Latn. Crúbadán language data for Eastern Dan uses: `dnj-x-east` [7]but it is unclear if that corpus is based on the same orthography as this one, even if it is from the same language variety.

Font Note: It has been Hugh's professional experience that in many cases fonts used to encode minority languages often fail to include two very important features. The first is that some classes of diacritics and characters do not combine elegantly for users. For instance: 〈◌̊〉 U+030A 'COMBINING RING ABOVE', does not elegantly combine with 〈🦄〉 U+1F984 'UNICORN FACE' to allow users to put a ring on the unicorn's horn‽ The second case is more grammatical in nature. Most fonts don't support 〈‽〉 U+203D 'INTERROBANG'.

Latin Orthography History

Orthography Note: It is the case that there are multiple writing systems for different speech varieties of the same ISO 639-3 designated languages, simultaneously. That is separate groups (socio-logical, or dialectical, or both) , are writing the same "language" in different ways at the same time.

Version	Date	Evolutionary steps	Mentor/Artist	Reference

Version 0.1	pre-1970 protestant	Imported from Liberia	Mission Biblique	R & V Forthcoming[8].
Version 0.2	pre-1970 catholic	concurrent with but separate from version 0.1	Roman Catholic Church	R & V Forthcoming[9].
Version 0.3	1974	??	Margrit Bolli / Eva Flik	Tiémoko Sébastien Bab[10] (reader; no orthography statement) R & Forthcoming[11]
Version 1	1982-1990	??	Margrit Bolli / Eva Flik	Bolli & Flik[12](Transitio Primer)
Version 2	1994	The start of using double U+0022 at the end of words appears in a course book for learning to read.		Bolli & Flik[13] (Transitional Primer)
Western Dan	2000	In *Western Dan* Biblical text preprints (for community circulation) use U+2013 instead of U+002D to indicate tone. (Forever muddling which character is correct in all future writing.)	Margrit Bolli / Eva Flik	See Ruth[14] an Jonah[15] Published in 2000.
		These texts contain U+201C,U+201D, and U+0022 as		

Version 3	(2005??)-2014	tone markers before and after words. (It might have been the idea that only U+0027 would be used twice and that human input habits chose to input U+0022 as a quicker step, and then word processing software auto-corrected some of these to U+201C, and U+201D)	Margrit Bolli/Valentin Vydrin	This corpus is representative this stage in th orthography.
Version 4	2014-2017+	There are significant changes to vowel and tone markers. In general away from digraphs towards single graphemes, and away from pre and post stem tone indication towards diacritic indication of tone.	Valentin Vydrin	Roberts, Brown Vydrin Forthcoming[16] & V Forthcoming[17] & R Forthcoming[18]

Corpus Description

Writing system

- BCP47: dnj-Latn
- Orthography version: 3

Writing System Note: In this writing system tone is shown in part through characters with the Unicode attributes for punctuation. Various characters before or after the stem (word) indicate the pitch melody of the orthographic word. These characters are not used in expected ways according to their Unicode attributes as encoded in the original documents for this corpus. As a result many applications do not properly type set or interact with the "words" in the ways that many users of "global" languages expect. One notable result is that the use of space around proper punctuation marks is not always as one would expect for an orthography written in a Latin script. That is, it is not uncommon to see something like `"ban- ? =Yaa'-` where there are extra spaces around the question mark.

Writing system, orthographic, linguistic, and alphabet descriptions for encoding of text in Eastern Dan version 3.

The closest thing to a formal writing system description for Eastern Dan is a 1994[19] community oriented reader which covers, Vowels, Consonants, Numbers, and punctuation. The 1994 reader improves upon a 1982 community oriented reader[20] by offering sections on numbers and punctuation. However, neither book presents an alphabetic order, or and alphabet in whole. Several forthcoming works do offer a formal linguistic description of the orthography, orthography testing, and a newly proposed orthography, but these works fail to provide details at the technical and writing system levels, focusing rather on the correspondences between linguistic units and typographical units.

In this section a short prose discussion is followed by a chart. Charts are followed by list presented in crucial ordering for tokenization by the python library segments.[21]
Note: the graphemes used here, with the exception of those recommended for special status by RFC3986[22] are presented because they are evidenced in the corpus.

These definitions and conventions are observed throughout this work:

- An **alphabet** is a list of **letters** used to transcribe a language. Alphabets usually have an order for pedagogical purposes, and for dictionary sorting purposes. At a technical level, SIL's NRSI[1] provides this: *a segmental writing system having symbols for individual sounds, rather than for syllables or morphemes. In a true alphabet, consonants and vowels are written as independent letters, in contrast to an abugida or an abjad. In a perfectly phonemic alphabet, phonemes and letters would be predictable in both directions; that is, the sound of a word could be predicted from its spelling and vice-versa. A phonetic alphabet is also predictable in this way, however it uses separate letters for separate allophones, whereas a phonemic alphabet may describe allophones of the same phoneme using a single letter*.
- **Letters** are typographical units for the purposes of pedagogy.
- **Characters** are single Unicode code points.
- **Graphemes** are typographical units. Often in a writing system these units carry meaning.

- **Multigraph** (from SIL's NRSI) a combination of two or more written symbols or orthographic characters (e.g. letters) that are used together within an orthography to represent a single sound. (Combinations consisting of two characters are also known as **digraphs**.).
- A **linguistic description** would include phonetic or phonological details for the characters used in the encoding of the text.
- A list of **phonemes** is a list of unique and distinctive sound units in a language. Many times an alphabet is based on a list of phonemes. But to the extent that two typographical characters are used together in a pattern (digraph) to indicate when co-occurring that they represent a phoneme then an alphabet might have fewer **letters**/components than a list of phonemes in the same language.
- A **writing system description** includes things like *casing correspondences*, *usage rules for casing*, *punctuation characters*, *usage rules for punctuation marks*, *letters*, *numbers*, and *characters used in Internet use*, with their Unicode code points used in technical encodings. A writing system description, more than just an orthography is needed to fully support a language on digital tools. It is necessary for creating a **Locale** description and is useful for creating a custom Keyboard layout, and other *Natural Language Processing Tools*.
- The following characters are used to provide special meaning to text outside of tables:
  - Content within square brackets denotes either phonetic representations or ISO639-3 codes `[]`.
  - Content within forward slashes denotes phonemic representations `//`.
  - Content within angle brackets orthographic or graphemic representations `〈〉`.
  - Content within double-slashes or pipes morphophonemic representations `// //` or `| |`.
  - In prose sections, Unicode characters will appear in the following order upon first mention: 〈⸘〉 U+203D 'INTERROBANG' a more natural prose style using one or more of the three referents will be used for following mentions.

Casing rules

Based on data within the corpus, casing rules appear to follow general French casing norms, with two noted exceptions.

1. Tone marks preceding the [a-zA-Z] portion of the word do not get capitalized, but the characters following the tone marks [a-zA-Z] do get capitalized.
2. The first word of a sentence is capitalized.
3. Proper nouns are capitalized.
4. Unlike French where, when an article is the first word of a sentence both the first word and the second word are capitalized, in Eastern Dan only the first word is capitalized.
5. Surnames are not capitalized as is the custom in French literature.
6. Uppercase can be used as a style choice in titles of creative works, much as is the case in many languages, which use a Latin script.

7. Only the first letter of a digraph is capitalized. i.e. 〈"Ɛa-〉 is correct whereas 〈"ƐA-〉 is not.

Punctuation

Based on data within the corpus, the following punctuation marks are observed. Their usages, as far as can be determined, from the corpus are indicated in the table.

Codepoint	Grapheme	Usage
U+00B0	°	Used as part of the abbreviation for number 〈n°〉.
U+005F	_	unknown
U+005B	[	unknown
U+005D	]	unknown
U+2026	…	unknown
U+201A	,	Errors - Should be U+002C
U+002F	/	unknown
U+00BB	»	Closes a direct speech statement
U+00AB	«	Opens a direct speech statement
U+0021	!	Closes an exclamation, interjection or emphatic statement
U+003B	;	unknown
U+2039	‹	Opens a quote inside of a direct speech statement
U+203A	›	Closes a quote inside of a direct speech statement
U+003C	<	Error - All cases are double i.e. << and should be replaced with U+00AB
U+003E	>	Error - All cases are double i.e. >> and should be replaced with U+00BB
U+003F	?	Closes a question statement
U+002E	.	unknown

U+002C	,	unknown
U+0029	)	Closes a parenthetical. Often a number, but sometimes a word in another language, or an alternate transcription of a name.
U+0028	(	Opens a parenthetical. Often a number, but sometimes a word in another language, or an alternate transcription of a name.
U+003A	:	unknown
U+002B	+	Precedes a telephone number to indicate country code

```
。
_
[
]
…
,
/
»
«
!
;
‹
›
<
 >
 ?
.
,
 )
 (
:
+
```

Number Characters

As evidenced in the corpus, when writing Eastern Dan with the Latin script the following numbers are used.

Codepoint	Grapheme
U+0030	0

U+0031	1
U+0032	2
U+0033	3
U+0034	4
U+0035	5
U+0036	6
U+0037	7
U+0038	8
U+0039	9

```
0
1
2
3
4
5
6
7
8
9
```

Number oriented notes:

- Thousands separator is 〈.〉 U+002E 'FULL STOP'.
- The is a shortened form of the word "number" in many transcription traditions. Unicode has a special character for this 〈№〉 U+2116 'NUMERO SIGN'. Typographical norm in Dan appear to follows French social practice, rather than best practice for encoding. This was evidenced only one time in the corpus and is the source of 〈°〉 U+00B0 'DEGREE SIGN', and likely deserves further investigation before strong claims are made about what method should be used in Eastern Dan writing. Wikipedia suggests that "the numero symbol is not in common use in France and does not appear on a standard AZERTY keyboard. Instead, the French Imprimerie nationale recommends the use of the form 〈no〉 (an 〈n〉 followed by a superscript lowercase 〈o〉). The plural form 〈nos〉 can also be used. In practice, the 〈o〉 is often replaced by the degree symbol 〈°〉, which is visually similar to the superscript 〈o〉 and is easily accessible on an AZERTY keyboard."[23]

Reasonable characters needed for Internet use

According to RFC 3986 [24]the following characters are needed for reasonable Internet use in the URL and URI syntax. In the Internet domain these characters can sometimes have a reserved meaning. Therefore they should be given appropriate consideration in all orthographies. So while their typographical function may or may not be present in the everyday writing of Eastern Dan, as Eastern Dan speakers become more digitally active with their language, these characters will increase in their usage by Eastern Dan language users.

This does not preclude any language based denotation that the orthography may make on these characters. For instance there is a long typographical history in Eastern Dan of using 〈=〉 U+003D 'EQUALS SIGN' as a tone marking character. It is even the case that the original text of this corpus was encoded with this character, no doubt for practical reasons of keyboard accessibility. However the more appropriate character is 〈꞊〉 U+A78A 'MODIFIER LETTER SHORT EQUALS SIGN'. Typographically across fonts, it is common that these characters appear the same, however their Unicode properties are different. U+A78A can not be substituted for Internet use and U+003D will not properly join with other text to form words in text processing software. By way of analogy, just because the internet does not use the same quote marks that French and Eastern Dan do does not mean that these languages need to change, only that accessing these characters and their social contribution is a needed consideration in orthography statements and written language development.

Unmentioned in RFC3986 is the use of 〈"〉 U+0022 'QUOTATION MARK', 〈>〉 U+003E 'GREATER-THAN SIGN', and 〈<\〉 u+003c="" 'greater-than="" sign'="" which="" are="" all="" highly="" important="" in="" various="" mark-ups="" like="" html525. Markdown[26], a common text markup language, requires 〈`〉 U+0060 'GRAVE ACCENT', 〈|〉 U+007C 'VERTICAL LINE', and 〈\〉 U+005C 'REVERSE SOLIDUS'. The following table represents RFC3986 plus 〈", <,>, |, `, \ 〉. Many of these characters are evidenced in the corpus. However some are not evidenced.

Codepoint	Grapheme
U+0021	!
U+0022	"
U+0023	#
U+0024	$
U+0025	%
U+0026	&

U+0027	'
U+0028	(
U+0029	)
U+002A	*
U+002B	+
U+002C	,
U+002D	-
U+002E	.
U+002F	/
U+003A	:
U+003B	;
U+003C	<
U+003D	=
U+003E	>
U+003F	?
U+0040	@
U+005C	\
U+005B	[
U+005D	]
U+005F	_
U+0060	`
U+007C	|
U+007E	~

```
%
:
/
?
#
[
]
@
!
$
&
'
(
)
*
+
"
,
;
=
-
.
_
~
"
`
|
>
<
```

Alphabet

Pedagogically the following as been presented in Eastern Dan "learning to write" materials[25].

Eastern Dan vowels carry distinctions for length, pitch, and nasality. Nasality is indicated by an ⟨n⟩following the vowel. Vowel length has been linguistically analyzed as two separate vowels and is indicated by sequential characters i.e. ⟨aa⟩. Some vowels are indicated by a digraph ⟨ɛa, aɔ⟩. These are not diphthongs (vowels that start at one phonetic value and finish at another value). Dieresis above vowels indicate a separate vowel quality. Vowels with dieresis are thought as a single character or letter of the alphabet. Dieresis is not a separable unit. The eng /ŋ/, orthographically indicated as ⟨ng⟩, is linguistically considered a vowel in Eastern Dan. This is in contrast to the typologically normal analysis and IPA symbol /ŋ/ usage as a consonant. Casing: for words starting with long/double vowels, only the first letter is case sensitive for sentence based casing rules. In this presentation of vowels, many vowels are presented, however, it is not true that this represents the Eastern Dan alphabet. The detailed representation here allows for vowels to be tokenized.

Aa aa

An an
Aan aan
Aɔ aɔ
Aɔn aɔn
Bh bh
Dh dh
Ee ee
Ɛ ɛ
Ɛɛ ɛɛ
Ɛn ɛn
Ɛɛn ɛɛn
Ë ë
Ëë ëë
Ën ën
Ëën ëën
Ɛa ɛa
Ɛan ɛan
Gw gw
In in
Iin iin
Ɩ ɩ
Ɩɩ ɩɩ
Kw kw
Ng ng
Oo oo
Ɔ ɔ
Ɔɔ ɔɔ
Ɔn ɔn
Ɔɔn ɔɔn
Ö ö
Öö öö
U u
Un un
Uun uun
Ü ü
Üü üü
Ün ün
Üün üün
Ÿ ü
Ÿü üü
Ʊ ʊ
Ʊʊ ʊʊ

Vowels

Phoneme chart (Oral)[SIL1982,V&K 2008,Ch10]

Linguistically, Eastern Dan is claimed to have a 12 point vowel system with length, pitch, and nasalization distinctions. Pitch patterns are covered under the tone marking section. Nasalization is occurs phonemically on 9 vowels. The velar nasal /ŋ/, orthographically indicated as 〈ng〉, is linguistically considered a vowel in Eastern Dan This bring the total to 22 vowels.

Oral	Front Unrounded	Back Unrounded	Back Rounded
Close	i	ɯ	u
Near-close			
Mid	e	ɤ	o
Open-mid	ɛ	ʌ	ɔ
Near-open	æ		
Open		a	ɒ

Nasal	Front Unrounded	Back Unrounded	Back Rounded
Close	ĩ	ɯ̃	ũ
Near-close			
Open-mid	ɛ̃	ʌ̃	ɔ̃
Near-open	æ̃		
Open		ã	ɒ̃

/ŋ/

Codepoint	Grapheme	IPA equivalent	Phonetic description
Uppercase, lowercase	,		
U+004E U+0067, U+006E U+0067	Ng, ng	ŋ	Velar Nasal

U+0041 U+0061 U+006E, U+0061 U+0061 U+006E	Aan, aan	ãã	long nasalized front open unrounded vowel
U+0041 U+0061, U+0061 U+0061	Aa, aa	aa	long front open unrounded vowel
U+0190 U+0061 U+006E, U+025B U+0061 U+006E	Ɛan, ɛan		
U+0190 U+0061, U+025B U+0061	Ɛa, ɛa		
U+0041 U+0254, U+0061 U+0254	Aɔn, aɔn		
U+0041 U+0254, U+0061 U+0254	Aɔ, aɔ		
U+0041 U+006E, U+0061 U+006E	An, an	ã	short nasalized front open unrounded vowel
U+0190, U+025B	Ɛ, ɛ	ɛ	
U+00CB, U+00EB	Ë, ë		
U+00D6, U+00F6	Ö, ö		
U+00DC, U+00FC	Ü, ü		
U+0045, U+0065	E, e		
U+0049, U+0069	I, i	i	
U+0186, U+0254	Ɔ, ɔ		
U+0041, U+0061	A, a		
U+004F, U+006F	O, o		
U+0055, U+0075	U, u		

ng

```
ɛa
ɛ
ë
ö
ü
e
i
aɔ
ɔ
a
o
u
```

Consonants

Phoneme chart[SIL1982,V&K 2008,Ch10]

	Labial	Dental	Palatal	Velar	Labio-velar
Voiceless Stops	p	t		k	kp, kw
Voiced Stops	b	d		g	gb, gw
Voiceless fricatives	f	s			
Voiced Fricatives	v	z			
Implosives	ɓ	ɗ			
Continuants		l	y		w

The presentation order of consonants here does not represent the alphabet of Dan, but rather the order required to tokenized the text into phonemes.

Codepoint	Grapheme	IPA equivalent	Phonetic description
Uppercase, lowercase	,		

```
kp
kw
k
gb
gw
g
bh
dh
```

```
m
n
f
s
v
z
l
w
r
y
```

Tone marking

Codepoint	Grapheme	IPA equivalent	Phonetic description	Usage Note
No Casing	,			

Pre-Stem

```
'
=
-
```

Post-Stem

```
-
'-
'
```

Unicode PUA reliance

Some texts have relied on Unicode PUA code points (U+E000..U+F8FF). All Dan texts, should be checked for PUA characters. Known used characters have been:

- Usage of U+F173 COMBINING MACRON-GRAVE. U+F173 was deprecated because the character was added to Unicode 5.0 as U+1DC6. There were 22 occurrences in a toolbox file which is not part of this corpus.

Content

This is about 20 issues of a 4 page monthly newsletter/newspaper published between 2005 and 2008. There are several chapters of *While waiting for a medical doctor*.
A new testament is also known to exist, but is not included in this repository or character counts.

Metrics

Pre text clean up stats

It should be noted that the percentages of characters and the percentages of phonemes presented here are attested only in this corpus. This corpus is not necessarily natural speech, and some characters may be over represented because *-Pamɛbhamɛ*, which was targeted at new readers, published a chart of the alphabet in nearly every issue, with some, but not many, words in French.

First round were off a bit because 4 issues of the local news paper did not get added to the file `mass-text.txt` (later renamed to `proof-of-concept-text.txt`), round three includes all the issues of *-Pamɛbhamɛ* and the chapters of *While waiting for a medical doctor*.
Linux Command Line:
`wc -l -w -m`

Round	Lines	Words	Characters
First	11686	46192	221389
Second	14491	55986	269437
Third	15756	86466	416782

UnicodeCharacterCount Stats for round three:
Presented in frequency order.

Provenance and text conditioning

Valentin Vydrin `vydrine[at]gmail[dot]com` Provided the corpus. Issues of the Eastern Dan local newpaper *-Pamɛbhamɛ* were provided as a series of `.doc` files. Three translated texts (trnaslated portions of *While waiting for a medical doctor*) were provided as a series of `.txt` files in related folders: moyan-sanni_ko_dhotroo, moyan-waa_won, moyan-yii_to_gu.

One `.doc` file was provided with 22 short (single paragraph length) parallel texts (Eastern Dan - French). And a copy of the New Testament was also provided but is not included in this corpus for copyright reasons.

Hugh Paterson III `sil.linguis[at]gmail[dot]com` converted the files following the steps in the `File types > Converted files` section.

File types and purpose

Original Files

`[gG]weta*.doc` these are the original files provided by VV.

`[gG]weta*.pdf` these are PDFs generated my MS Word by Rebecca Paterson from files provided by VV.

`[gG]weta*.txt` these files are generated by Hugh Paterson using `pdftotext`.

`*-sfm.txt` files have a hand coded structure to them that includes making for things like newspaper title, volume, date, tagline, article, heading 1, heading 2, and text of article:

```
\newspaper -Pamɛbhamɛ
\volume-eng 001
\volume-or "Nimlɒɒ : 00x---
\date 2005 'Zë Zë -kwɛ
\tagline "su –bha ‘sëëdhɛ -mü "Gwɛɛtaawo
\body
\article 1
\heading 1
\heading 2
\p 1
```

Three folders containing some `.txt` files are held in the `While-waiting-for-a-medical-doctor` directory.

- moyan-sanni_ko_dhotroo
- moyan-waa_won
- moyan-yii_to_gu

The folder `sil-pua` contains `teckit` files for transferring deprecated Unicode points from SIL's PUA area to their accepted and final Unicode point values.

Converted Files

The following transforms were performed on the original files to extract the text from the originally provided formats, and to clean up character inconsistencies, so that corpus analysis

for text input could be optimized.

The issues of *-Pamɛbhamɛ* (provided as `[gG]weta*.doc`) were converted to PDFs by opening them in Microsoft Word 16.13.1 (180523) on MacOS 10.13.3. The operating system Print option was invoked, and the "Save as PDF" option was used. The PDFs were transfered to an Ubuntu machine where `pdftotext` was used to extract the text to `.txt` files. The multitude of text files were then concatenated to a single file `mass-text.txt` using the following commands on Ubuntu 16.04 (`$` represents the start of the command line, and the command was executed from the root of this repo):

- `$ cp $( find ./*Pam*/*weta*/*weta*.pdf ) . && for f in *weta*.pdf; do pdftotext $f mass-text_$f.txt; done && rm *.pdf && cat mass-text*.txt >> combined-gweta-text.txt && rm mass-text_*.txt`

Each of the three sets of files in the directory `While-waiting-for-a-medical-doctor` were concatenated together with the following:

- `$ cp $( find ./While-waiting-for-a-medical-doctor/*moyan-*/*moyan-*.old.txt ) . && cat moyan-sanni*.old.txt >> combined-moyan-sanni_ko_dhotroo.old.txt && cat moyan-yii*.old.txt >> combined-moyan-yii_gu.old.txt && cat moyan-waa*.old.txt >> combined-moyan-waa_won.old.txt && rm moyan-*.old.txt`

These files were then visually inspected in the text editor Atom prior to further processing. Upon visual inspection HTML style heading tags `<h>` and `</h>` were noticed.

The combined issues of *-Pamɛbhamɛ* and the three files representing *While waiting for a medical doctor* were then concatenated into the same file for character level processing.

- `$ cat combined-*.txt >> proof-of-concept-text.txt && rm combined-*.txt`

Character Maintenance

1. Teckit was used to make sure that all deprecated PUA Unicode code points moved to current (Unicode 10) code points.

```
$ txtconv -i proof-of-concept-text.txt -o proof-no-PUA.txt -t sil-pua/SILPUA.tec -i
```

2. Remove all BOM marks (they were created or concatenated into the middle of the file with the `cat` command).

```
$ cat proof-no-PUA.txt | perl -CS -pe 's/\N{U+FEFF}//g' > proof-no-PUA-no-BOM.txt
```

```
3. Markup tags were removed from the text with search and replace. `<h>` and `</h>`


#### Typographical Encoding Errors
In the course of text production several different look-alike characters have been

1. Correct equal signs

Replace normal equal sign U+003D with letter equal sign U+A78A.
```

```
cat proof-no-PUA-no-BOM-no-TAGS.txt | perl -CS -pe 's/\N{U+003D}/\N{U+A78A}/g' > Co
```

2. Replace U+FFF9 with 'LATIN SMALL LETTER U WITH GRAVE' (U+00F9) target 34

```
Corrected-equal.txt | perl -CS -pe 's/\N{U+FFF9}/\N{U+00F9}/g' > Corrected-equal-letterU.txt
```

```
3. Corrected bad non-breaking hyphen.
```

Corrected-equal-letterU.txt| perl -CS -pe 's/\N{U+001E}/\N{U+02D7}/g' > Corrected-equal-letterU-nbs.txt

```
4. Corrected bad commas U+201A --> U+002C
```

Corrected-equal-letterU.txt| perl -CS -pe 's/\N{U+001E}/\N{U+02D7}/g' > Corrected-equal-letterU-nbs.txt

```
5. replace Non-breaing space U+00A0 with normal space U+0020 target 374


U+0009        482
U+000A        30690
U+000C        220
U+000D        1340
U+001E        5442
U+0020        124711
```

6. Correct **minus** signs
 Underscore, dash, **and minus are** all moved **to** U+02D7 which **is** modifier letter minus

  ```
  sed 's/[_ –-]/$(echo -ne '\u02D7')/g' mass-text.txt > spell-corrected-mass-text.t
  ```

 This solution **is** too greedy. I need **to convert** hyphens **between** numbers back **to** hyp

7. Corrected non-letter apostrophe **to** letter apostrophe

8. Correct **double** apostrophe **to** proper quote marks.
9. French Quotes

Bibliography

<!-- 1 Footnote content here. [↩](#a1)
<b **id**="f2">2 Footnote **content** here. [↩](#a2) -->

<b **id**="f3">3 Simons, Gary. F., & Charles D. Fennig (Eds.) 2017. Ethnologue: Lan

<b **id**="f4">4 Roberts, David & Valentin Vydrin. Forthcoming. Chapter 10: Easter

<b **id**="f5">5 Valentin Vydrin. 2012. ISO 639-3 **Change** Request 2012-083. **Online**: h

<b **id**="f6">6 Phillips, A. & M. Davis (Eds.) 2009. Tags **for** Identifying Language

<b **id**="f7">7 Scannell, Kevin (Ed.) 2009. An Crúbadán - Dan. Saint Louis Univer

<b **id**="f8">8 Roberts, David & Valentin Vydrin. Forthcoming. Chapter 10: Easter

<b **id**="f9">9 Roberts, David & Valentin Vydrin. Forthcoming. Chapter 10: Easter

<b **id**="f10">10 Baba, Tiémoko Sébastien .1978. Yaobhaa -wo bhe pe -se -ya ’gu (R

<b **id**="f11">11 Roberts, David & Valentin Vydrin. Forthcoming. Chapter 10: Easte

<b **id**="f12">12 Bolli, Margrit & Eva Flik. 1982. Guide d’orthographe pour la lan

<b **id**="f13">13 Bolli, Margrit & Eva Flik. 1994. Cours-eclair de lecture pour de

<b **id**="f14">14 Bolli, Margrit & Eva Flik. 2000. Rutö. Société Internationale de

<b **id**="f15">15 Bolli, Margrit & Eva Flik. 2000. Zonasö. Société Internationale

<b **id**="f16">16 Roberts, David, Dana Basnight-Brown & Valentin Vydrin. Marking t

<b **id**="f17">17 Roberts, David & Valentin Vydrin. Forthcoming. Chapter 10: Easte

<b **id**="f18">18 Vydrin,Valentin & David Roberts. Forthcoming. Tonal oral reading

<b **id**="f19">19 Bolli, Margrit & Eva Flik. 1994. Cours-eclair de lecture pour de

<b **id**="f20">20 Bolli, Margrit & Eva Flik. 1982. Guide d’orthographe pour la lan

```
<b id="f21">21 </b>Moran, Steven & Robert Forkel. 2017 (November 16). cldf/segments

<!--
```

Some text [1]

```
Then from within the footnote, link back to it.
```

1 Footnote content here. ↩
```

-->

# Intellectual property ownership and licenses

## Text (corpus) content

Copyright claims are un-clear.
If authors of content were employed by SIL, SIL International would be the copyright owner. (This is only relevant because the works themselves do not have copyright claims or licenses attached, but do reference SIL's address.) Otherwise copyright belongs to the authors, or their employer. It does not readily seem that the authors are attributed in the corpus, but they might be in the orthography.

Only copyright owners can license materials. Therefore this content bears no license, as Hugh makes no content claims on the content of the corpus, and did not receive content under license. Use under the *fair use* doctrine is assumed.

## Hugh Paterson's Contribution

The `README.md` which is Hugh Paterson III's contribution is copyright Hugh Paterson III 2018, and licensed under the Creative Commons Attribution 4.0 License.

The `generate-corpus.bash` script is also Hugh's contribution and is licensed under the MIT version provided.

## SIL International's Contribution
```

Other content such as the content contained under the folder `/SILPUA` is licensed as originally offered (MIT).