

# Creating a Corpus and Chained Bigrams for Spanish Keyboard Development and Evaluation

Ian Douglas, B.Sc

[ian@zti.co.za](mailto:ian@zti.co.za)

5 September 2021

Version 1.0.1

DOI: <https://doi.org/10.5281/zenodo.5501913>

This work is licensed under the Creative Commons Attribution 4.0 International License.

Unqid: f6fa237bd77f7d247ce264a0b575121f

Please check via DOI for latest version.

## Abstract

The process to create a corpus suitable for evaluating computer keyboard layouts optimised for typing Spanish. After sourcing, sampling and cleaning suitable texts, the texts are processed to extract bigrams, which are then used to create sample input texts of a desired length. These texts have a character distribution, and letter sequence, closely matching Spanish, even though they look random. The resulting texts are excellent for evaluating keyboard layouts. Corpus analysis is included.

**Keywords:** Spanish text corpus, Spanish letter frequency, bigram frequency, Spanish keyboard layout, Spanish keyboard layout evaluation.

## Contents

1. Introduction
2. Creating the Spanish corpus
3. Corpus analysis
4. Creating chained bigrams (Markov chains) and texts
5. List of datasets and files
6. Acknowledgements
7. Bibliography

## Revision history:

12 September 2021 1.0.0 Initial version.

## 1. Introduction

This project was done as part of the process to develop a Spanish-language version [1] of Arno Klein's Engram keyboard layout [2]. The project was initiated by Nick Gutiérrez [3], with Miguel Guzmán [4] also contributing.

When designing or evaluating a computer keyboard layout for a given language, it is necessary to know the character frequency for that language. It is also useful to know the bigram and trigram frequencies. These frequencies are calculated by analysing a suitable corpus of text.

However, the available corpora, or indeed analysis, was driven by other needs, typically cryptographic or lexical analysis, which are totally different to the keyboard layout problem. So I compiled a new corpus.

The corpus collection was done during July and August 2021, all files sourced from the Internet are as of that date.

## 2. Creating the Spanish corpus

There is no freely-available Spanish corpus or analysis that is suitable for keyboard analysis. There is an existing letter frequency list at sttmedia.com [5] as well as samples of n-grams at ngrams.info [6].

I followed a similar approach to how I did the English corpus [7], and created a suitable corpus, from two sources: assorted web texts from the Leipzig Corpus Collection [8], [9], and extracts from Spanish books provided by Project Gutenberg [10].

For Leipzig, I downloaded the largest version of each file, except those marked as "non-Spain", and extracted the "Sentences" files. These were all merged into one file of 3.4 GB for cleaning.

The Leipzig collection is scraped from the web, sentences extracted, numbered, and put in a file. Leipzig does some analysis, but it was not relevant for this exercise. Their sources included news sites, which reference many non-Spanish places, people and brands. Also, using the Web as a source means that there is also non-standard language, and perhaps more technical strings like "https://" or foreign characters than the average person types.

So to clean the file, I needed to:

1. Remove the line numbers
2. Remove any characters that were not in the proposed character set.
3. Remove as many non-Spanish words as possible.

The defined character set, after some discussion, was set at

a b c d e f g h i j k l m n o p q r s t u v w x y z  
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z  
á é í ó ú ü ñ ç Á É Í Ó Ú Ü Ñ Ç  
1 2 3 4 5 6 7 8 9 0  
? ¿ ¡ ! @ # \$ % ^ & \*  
~ \_ - — = + \ | /

( ) [ ] { } < >  
" " ' ' " " .. ^ ^  
.. ; ; .  
o a €  
space

This adds some characters not on the current standard Spanish keyboard, and drops some that are. We discussed whether to do the diacritics via “dead keys” or to put them on the third level to be accessed via the AltGr key. Arno decided to go with the “dead key” method, while Nick and myself created layouts that use the AltGr method. Depending on the operating system and layout driver, it may be possible to use dead keys, AltGr, and the Linux “Compose” function to create more varied diacritics (e.g. for French or German) quite easily.

The clean-up process does a line at a time by first removing the line numbers, then deleting all characters in the required list. If any characters remained, the line was discarded, otherwise the original line went to the next step.

We then tried to eliminate lines with non-Spanish words. This was the most difficult part, and eventually hit the point of diminishing returns. Since much of the Leipzig corpus was sourced from news sites, there were many foreign place, personal, and other nouns. To strip out non-Spanish words, I needed a Spanish word list. I tried the Linux word list, but that did not include all tense, gender or plural variations. I eventually downloaded the word lists from ListaPalabras [11].

I then matched each word in the corpus against these words, including dropping terminal “s” or “es” to find matches. Words not found were written to a file, and the line temporarily dropped. Miguel went through the most frequent unknown words, and either approved, rejected or corrected them. These changes were then incorporated into the next run. After a few of these cycles, we decided that we were getting diminishing returns, so any lines with the remaining unknown words were discarded.

The initial file had :

26,329,994 lines, (one sentence per line)

591,648,169 words, and

3,648,255,303 characters.

Since natural text is normally in paragraphs rather than lines, I artificially split the text into paragraphs, according to this method:

- If the word count was 200 or more, or sentences were 8, end the paragraph.
  - If the sentences were 7, 90% chance of ending paragraph.
  - If the sentences were 6, 80% chance of ending paragraph.
  - If the sentences were 5, 70% chance of ending paragraph.
  - If the sentences were 4, 60% chance of ending paragraph.

- If the sentences were 3, 50% chance of ending paragraph.

Otherwise continue building a paragraph.

The final “cleaned” text had:

5,480,692 lines, (one paragraph per line)

265,281,891 words, and

1,610,201,701 characters

From this I was able to extract character and bigram counts, the results are below. The cleaned file is posted in the accompanying data set on Zenodo.

Since the style of writing on the web is different to that used in conventional books, I followed a similar process to what I did for the English corpus. I used the tools provided by Martin Gerlach and Francesc Font-Clos [12] to get Spanish-language books from Project Gutenberg, and took extracts. For each book, if the word count was over 10,000, I would skip the first 200 lines (Gutenberg front matter and contents), and then take an extract which met the character set requirements. Of the 702 books, 517 allowed extracts of around 10kB, and of those, 473 allowed extracts of around 20kB. This produced 517 files ranging from 10 to 28 kB in size, which were concatenated into one file of 10.8 MB. This file has:

234,541 lines,

1,859,604 words, and

11,275,853 characters.

### 3. Corpora analysis

The resulting files were analysed for letter frequency, words, and bigrams. Arno indicated that he did not need higher n-grams for his layout generation algorithm.

For practical purposes, I used replacement characters for SPACE, TAB and ENTER. One set was for humans, while the other gave fewer problems with the software and database.

Character	ASCII decimal	Unicode	For Humans	For computers
Space	32	U+0020	␣	§
Tab	09	U+0009	→	↔
Enter	13	U+000D	↔	¶

Table 1: Replacement characters used

The final character frequency for the Leipzig corpus is in Table 2.

Character	Count	Percentage	Character	Count	Percentage
␣	262541545	16.65452	O	800399	0.05077

Creating a Corpus and Chained Bigrams for Spanish Keyboard Development and Evaluation  
I Douglas 2021

<i>Character</i>	<i>Count</i>	<i>Percentage</i>	<i>Character</i>	<i>Count</i>	<i>Percentage</i>
e	160935194	10.20904	-	787681	0.04997
a	146340981	9.28325	4	785123	0.04980
o	102702819	6.51503	8	766345	0.04861
n	88162627	5.59266	k	737458	0.04678
s	82333504	5.22289	6	707311	0.04487
r	82239923	5.21695	7	685187	0.04347
i	78642516	4.98875	Y	510288	0.03237
l	68366313	4.33687	?	484780	0.03075
d	62095764	3.93909	w	342214	0.02171
t	54350794	3.44778	'	313868	0.01991
c	52950737	3.35897	¿	304010	0.01929
u	49852653	3.16244	%	292908	0.01858
m	32440338	2.05788	K	244584	0.01552
p	31329186	1.98739	W	242247	0.01537
,	15026080	0.95319	Q	241397	0.01531
b	14600048	0.92616	;	236503	0.01500
.	14035109	0.89033	»	204119	0.01295
g	13778614	0.87406	«	200707	0.01273
v	11791362	0.74799	Z	139001	0.00882
q	10963958	0.69551	X	138406	0.00878
ó	10647166	0.67541	!	104179	0.00661
y	9927001	0.62973	/	95460	0.00606
h	9328700	0.59177	\$	79889	0.00507
f	8941976	0.56724	Á	75270	0.00477
í	5972170	0.37885	,	71710	0.00455
á	5464744	0.34666	í	62125	0.00394
E	5313264	0.33705	‘	58537	0.00371
j	5068794	0.32154	É	42845	0.00272
z	4761007	0.30202	º	40589	0.00257
C	3582383	0.22725	ü	40265	0.00255
A	3380251	0.21443	*	39979	0.00254
é	3249282	0.20612	—	33017	0.00209
S	3101351	0.19674	&	18889	0.00120
0	3055962	0.19386	‘	16440	0.00104
P	3031108	0.19228	+	13789	0.00087
L	3014046	0.19120	ç	13660	0.00087
1	2572934	0.16322	ª	12391	0.00079
M	2517620	0.15971	Ó	12372	0.00078

<i>Character</i>	<i>Count</i>	<i>Percentage</i>	<i>Character</i>	<i>Count</i>	<i>Percentage</i>
"	2508419	0.15912	@	11420	0.00072
ñ	2149556	0.13636	'	9808	0.00062
x	2085747	0.13231	Í	9572	0.00061
2	2007512	0.12735	.	8896	0.00056
D	1712540	0.10864	Ú	8411	0.00053
ú	1598899	0.10143	€	4216	0.00027
T	1587441	0.10070	>	4015	0.00025
R	1445114	0.09167	Ñ	3126	0.00020
N	1441953	0.09147	<	1145	0.00007
B	1359070	0.08621	¬	537	0.00003
I	1253220	0.07950	~	536	0.00003
F	1205344	0.07646	..	391	0.00002
G	1185853	0.07523	Ü	197	0.00001
9	1128483	0.07159	Ç	131	0.00001
U	1009741	0.06405	\	9	0.00000
3	942218	0.05977	{	0	0.00000
5	929391	0.05896	[	0	0.00000
H	925896	0.05873	}	0	0.00000
(	909292	0.05768	]	0	0.00000
)	888113	0.05634		0	0.00000
J	850601	0.05396	=	0	0.00000
V	836857	0.05309	—	0	0.00000
"	833074	0.05285	^	0	0.00000
"	832179	0.05279	#	0	0.00000
:	822983	0.05221			

Table 2: Character count and percentage in the Leipzig corpus

The 200 most common words in the Leipzig corpus (case-specific) are in Table 3.

<i>Rank</i>	<i>Word</i>	<i>Rank</i>	<i>Word</i>	<i>Rank</i>	<i>Word</i>
1	de	68	todos	135	nuevo
2	la	69	No	136	antes
3	a	70	Se	137	esa
4	que	71	porque	138	les
5	el	72	c	139	tico
6	en	73	hab	140	v
7	n	74	vez	141	embargo
8	y	75	Las	142	da

Rank	Word	Rank	Word	Rank	Word
9	los	76	me	143	ahora
10	se	77	t	144	estos
11	del	78	uno	145	gobierno
12	un	79	pol	146	hecho
13	con	80	cada	147	equipo
14	las	81	Es	148	Si
15	por	82	era	149	r
16	una	83	e	150	cuenta
17	para	84	mayor	151	tienen
18	su	85	dijo	152	mucho
19	es	86	M	153	ellos
20	no	87	personas	154	Un
21	al	88	Y	155	Sin
22	m	89	durante	156	aunque
23	El	90	nos	157	Adem
24	o	91	tica	158	siempre
25	lo	92	tres	159	Al
26	como	93	hace	160	seg
27	La	94	millones	161	tras
28	En	95	otros	162	tener
29	ha	96	tiempo	163	solo
30	os	97	contra	164	pueden
31	sus	98	ese	165	qu
32	fue	99	mismo	166	ver
33	est	100	sido	167	Este
34	ser	101	pr	168	otras
35	este	102	De	169	otra
36	as	103	vida	170	Gobierno
37	d	104	despu	171	Con
38	pero	105	Pero	172	horas
39	le	106	gran	173	Una
40	Los	107	hacer	174	mientras
41	esta	108	mi	175	San
42	son	109	fueron	176	grupo
43	sobre	110	primera	177	unos
44	entre	111	lugar	178	Su
45	dos	112	quien	179	manera
46	tambi	113	ciudad	180	Tambi

<i>Rank</i>	<i>Word</i>	<i>Rank</i>	<i>Word</i>	<i>Rank</i>	<i>Word</i>
47	ya	114	forma	181	Espa
48	l	115	bien	182	poco
49	desde	116	primer	183	va
50	an	117	mundo	184	poder
51	han	118	tanto	185	algunos
52	muy	119	menos	186	partido
53	todo	120	presidente	187	ten
54	hasta	121	eso	188	Nacional
55	tiene	122	otro	189	cuatro
56	pa	123	trabajo	190	esto
57	si	124	C	191	cual
58	sin	125	podr	192	adem
59	A	126	Para	193	debe
60	cuando	127	caso	194	final
61	parte	128	mejor	195	estado
62	p	129	momento	196	todas
63	donde	130	estar	197	mil
64	hay	131	ante	198	Estados
65	ni	132	pasado	199	muchos
66	Por	133	f	200	gente
67	puede	134	hoy		

Table 3: The 200 most common words in the Leipzig corpus.

The 100 most frequent bigrams, case sensitive, in the Leipzig corpus are in Table 4.

<i>Rank</i>	<i>Bigram</i>	<i>Rank</i>	<i>Bigram</i>	<i>Rank</i>	<i>Bigram</i>
1	a_	35	r_	69	ia
2	e_	36	al	70	si
3	o_	37	do	71	me
4	de	38	ad	72	ec
5	_d	39	qu	73	ma
6	s_	40	st	74	di
7	_e	41	_	75	ío
8	n_	42	_m	76	pr
9	en	43	na	77	nd
10	es	44	un	78	ne
11	_l	45	se	79	_h

Rank	Bigram	Rank	Bigram	Rank	Bigram
12	p	46	ro	80	ón
13	l	47	to	81	n
14	la	48	q	82	li
15	er	49	ca	83	is
16	c	50	in	84	r
17	a	51	da	85	mo
18	ar	52	lo	86	nc
19	ue	53	ic	87	pe
20	ra	54	ri	88	mi
21	el	55	ti	89	so
22	s	56	t	90	om
23	os	57	y	91	am
24	z	58	ac	92	j
25	re	59	po	93	su
26	nt	60	ie	94	ha
27	on	61	pa	95	sa
28	ci	62	tr	96	em
29	co	63	no	97	f
30	ta	64	io	98	ni
31	te	65	le	99	ce
32	an	66	y	100	E
33	as	67	u		
34	or	68	id		

Table 4: The 100 most frequent bigrams in the Leipzig corpus.

For the Gutenberg extracts, the character frequency is in Table 5:

Character	Count	Percentage	Character	Count	Percentage
u	1796755	16.29367	F	3939	0.03572
e	1077034	9.76696	Q	3667	0.03325
a	1018474	9.23591	2	2889	0.02620
o	741684	6.72588	»	2581	0.02341
s	647739	5.87395	)	2408	0.02184
n	567973	5.15060	«	2384	0.02162
r	547656	4.96636	(	2378	0.02156
l	475477	4.31181	[	2327	0.02110
i	473889	4.29741	]	2327	0.02110

Creating a Corpus and Chained Bigrams for Spanish Keyboard Development and Evaluation  
I Douglas 2021

<i>Character</i>	<i>Count</i>	<i>Percentage</i>	<i>Character</i>	<i>Count</i>	<i>Percentage</i>
d	418312	3.79341	3	2197	0.01992
u	353343	3.20425	5	2151	0.01951
t	345682	3.13478	8	2027	0.01838
c	326606	2.96179	w	2011	0.01824
↔	234541	2.12691	4	1940	0.01759
m	232693	2.11015	—	1654	0.01500
p	194795	1.76648	"	1591	0.01443
,	153722	1.39401	6	1516	0.01375
b	129075	1.17050	7	1506	0.01366
.	101198	0.91770	9	1252	0.01135
q	92314	0.83714	X	1245	0.01129
g	92208	0.83618	k	1182	0.01072
y	91133	0.82643	*	1155	0.01047
v	85807	0.77813	Á	868	0.00787
h	85653	0.77673	=	861	0.00781
í	56397	0.51143	Z	757	0.00686
á	53871	0.48852	É	738	0.00669
f	53739	0.48733	W	398	0.00361
ó	49627	0.45004	Í	366	0.00332
-	39035	0.35398	ü	365	0.00331
j	38157	0.34602	Ñ	295	0.00268
z	34851	0.31604	Ó	288	0.00261
é	31906	0.28934	K	269	0.00244
E	24167	0.21916	'	248	0.00225
A	21288	0.19305	"	222	0.00201
ñ	20566	0.18650	a	219	0.00199
-	17007	0.15423	"	210	0.00190
C	16372	0.14847	ç	201	0.00182
;	15903	0.14421	~	167	0.00151
S	15593	0.14140	º	156	0.00141
L	15413	0.13977	.	150	0.00136
P	14701	0.13331	Ú	62	0.00056
D	12536	0.11368	'	54	0.00049
M	12267	0.11124	`	53	0.00048
N	10646	0.09654	}	44	0.00040
I	10207	0.09256	,	41	0.00037
x	9702	0.08798	&	38	0.00034
:	8898	0.08069	#	34	0.00031

<i>Character</i>	<i>Count</i>	<i>Percentage</i>	<i>Character</i>	<i>Count</i>	<i>Percentage</i>
ú	8716	0.07904	{	32	0.00029
R	8626	0.07822	/	27	0.00024
!	7907	0.07170	Ç	15	0.00014
T	7755	0.07033	>	10	0.00009
i	7575	0.06869		10	0.00009
O	7494	0.06796	^	9	0.00008
?	7139	0.06474	<	9	0.00008
¿	6871	0.06231	+	8	0.00007
Y	6741	0.06113	\$	5	0.00005
V	5788	0.05249	'	2	0.00002
1	5284	0.04792	\	1	0.00001
B	5276	0.04784	..	1	0.00001
0	4900	0.04444	Ü	1	0.00001
H	4465	0.04049	@	0	0.00000
G	4403	0.03993	%	0	0.00000
U	4264	0.03867	¬	0	0.00000
J	3945	0.03577	€	0	0.00000

Table 5: Character count and percentage in the Gutenberg extracts

The 200 most common words in the Gutenberg extracts (case-specific) are in Table 6.

<i>Rank</i>	<i>Word</i>	<i>Rank</i>	<i>Word</i>	<i>Rank</i>	<i>Word</i>
1	de	68	c	135	nada
2	y	69	hombre	136	hacer
3	la	70	ella	137	mundo
4	que	71	usted	138	e
5	a	72	mismo	139	De
6	el	73	casa	140	fin
7	en	74	donde	141	do
8	los	75	tiempo	142	of
9	n	76	or	143	hombres
10	se	77	son	144	tres
11	las	78	te	145	tierra
12	con	79	ndose	146	puede
13	no	80	D	147	mo
14	un	81	otro	148	r
15	su	82	he	149	ese

Rank	Word	Rank	Word	Rank	Word
16	del	83	ten	150	is
17	por	84	hay	151	mujer
18	una	85	Pero	152	decir
19	lo	86	vez	153	Yo
20	m	87	fu	154	mano
21	al	88	all	155	aunque
22	para	89	poco	156	veces
23	como	90	aqueل	157	noche
24	es	91	esto	158	hace
25	sus	92	p	159	cada
26	le	93	A	160	cosas
27	o	94	the	161	antes
28	me	95	todas	162	espa
29	mi	96	siempre	163	medio
30	l	97	ojos	164	algo
31	as	98	Los	165	Espa
32	hab	99	gran	166	sido
33	El	100	desde	167	hecho
34	an	101	Se	168	esa
35	ni	102	otra	169	Si
36	era	103	estaba	170	S
37	si	104	han	171	f
38	sin	105	Dios	172	mayor
39	Y	106	pues	173	Es
40	pero	107	ver	174	entonces
41	os	108	despu	175	dar
42	La	109	uno	176	estas
43	ser	110	aqu	177	amor
44	ha	111	aquella	178	pueblo
45	todo	112	les	179	otras
46	yo	113	sino	180	ora
47	d	114	dijo	181	Las
48	tan	115	tu	182	pa
49	sobre	116	mas	183	voz
50	muy	117	cual	184	eran
51	cuando	118	mucho	185	cosa
52	dos	119	quien	186	dicho
53	ya	120	--	187	eso

Rank	Word	Rank	Word	Rank	Word
54	En	121	toda	188	Al
55	No	122	Por	189	cabeza
56	est	123	otros	190	mal
57	este	124	mis	191	estos
58	porque	125	menos	192	mejor
59	todos	126	tiene	193	ahora
60	bien	127	parte	194	alma
61	esta	128	Qu	195	algunos
62	hasta	129	tal	196	misma
63	qu	130	tanto	197	hizo
64	entre	131	tambi	198	madre
65	vida	132	don	199	tica
66	nos	133	padre	200	Juan
67	t	134	ellos		

Table 6: The 200 most common words in the Gutenberg extracts.

The 100 most frequent bigrams, case sensitive, in the Gutenberg extracts are in Table 7.

Rank	Bigram	Rank	Bigram	Rank	Bigram
1	e_	35	_m	69	_n
2	a_	36	te	70	_u
3	s_	37	or	71	mo
4	o_	38	al	72	sa
5	de	39	ad	73	di
6	_d	40	lo	74	ti
7	en	41	st	75	si
8	_e	42	se	76	ia
9	es	43	_y	77	su
10	n_	44	_q	78	io
11	_l	45	y_	79	ha
12	os	46	ci	80	_u
13	_	47	r_	81	ll
14	la	48	ro	82	_v
15	er	49	to	83	ec
16	ue	50	_t	84	ba
17	_c	51	da	85	ac
18	_s	52	ca	86	id
19	as	53	le	87	s,

<i>Rank</i>	<i>Bigram</i>	<i>Rank</i>	<i>Bigram</i>	<i>Rank</i>	<i>Bigram</i>
20	ra	54	ie	88	so
21	ap	55	no	89	mi
22	u	56	un	90	o,
23	a	57	in	91	.e
24	an	58	na	92	om
25	ar	59	ri	93	is
26	re	60	nd	94	a,
27	el	61	h	95	ic
28	qu	62	ab	96	am
29	do	63	e-e	97	pe
30	on	64	tr	98	ce
31	nt	65	me	99	ía
32	l	66	pa	100	r
33	co	67	po		
34	ta	68	ma		

Table 7: The 100 most frequent bigrams in the Gutenberg extracts.

Finally, I merged the two sets of character counts and bigrams, case- and diacritic-insensitively, to create datasets for Arno's programs. Other computer-based algorithms may require case and diacritic letters to be distinct.

<i>Character</i>	<i>Count</i>	<i>Percent</i>
u	262541545	16.65
E	169540585	10.75
A	155261246	9.85
O	114162756	7.24
N	91757262	5.82
I	85877478	5.45
S	85434855	5.42
R	83685037	5.31
L	71380359	4.53
D	63808304	4.05
C	56546911	3.59
T	55938235	3.55
U	52510166	3.33
M	34957958	2.22
P	34360294	2.18

<i>Character</i>	<i>Count</i>	<i>Percent</i>
B	15959118	1.01
G	14964467	0.95
V	12628219	0.8
Q	11205355	0.71
Y	10437289	0.66
H	10254596	0.65
F	10147320	0.64
J	5919395	0.38
Z	4900008	0.31
X	2224153	0.14
K	982042	0.06
W	584461	0.04

Table 8: Merged uncase, diacritic-striped Spanish character frequency.

The top 100 uncase, diacritic-striped bigrams are in table 9.

<i>Rank</i>	<i>Bigram</i>	<i>Rank</i>	<i>Bigram</i>	<i>Rank</i>	<i>Bigram</i>
1	DE	35	CA	69	NC
2	EN	36	RI	70	BA
3	ES	37	IN	71	NI
4	LA	38	MA	72	LL
5	ER	39	DA	73	CU
6	OS	40	IE	74	VI
7	ON	41	LE	75	EM
8	RA	42	PA	76	VE
9	UE	43	PO	77	OL
10	AS	44	TI	78	IT
11	AN	45	TR	79	RT
12	EL	46	SI	80	ED
13	AR	47	ME	81	AT
14	RE	48	DI	82	IR
15	CO	49	ND	83	IM
16	NT	50	IC	84	BR
17	DO	51	AC	85	BI
18	TA	52	MO	86	GU
19	QU	53	ID	87	MP
20	TE	54	EC	88	IL
21	CI	55	SA	89	US

Rank	Bigram	Rank	Bigram	Rank	Bigram
22	OR	56	AB	90	CH
23	AL	57	IS	91	EG
24	AD	58	SO	92	GA
25	IO	59	MI	93	OC
26	ST	60	SU	94	UR
27	SE	61	NE	95	UA
28	RO	62	PR	96	TU
29	LO	63	PE	97	UI
30	TO	64	HA	98	PU
31	NA	65	LI	99	VA
32	IA	66	OM	100	GO
33	NO	67	AM		
34	UN	68	CE		

Drawing 9: Most frequent uncase, diacritic-striped Spanish bigrams.

#### 4. Creating chained bigrams (Markov chains) and texts

Using the bigram counts, we can create bigram chains that Cervantes' Monkeys can use to create texts of arbitrary length, with a character and bigram frequency closely matching Spanish.

The procedure is as follows. There is a fuller discussion in my English corpus paper [7]

1. Decide on the required number of characters, for example 10,000. Add some excess capacity, say 10%.
2. Read in the bigram counts.
3. Add up the total number of bigrams,
4. Divide the number required, by the total. This gives us a scaling factor.
5. For each bigram, populate a table with (scaling factor × count) many bigrams. This creates a potentially large table.
6. When all bigrams are stored, shuffle the table.
7. Build an output text, starting with the first bigram.
8. Look at the second letter of this bigram, then search from where you are in table for the first bigram starting with this character. Add the second character of this bigram to the output, and loop this process until you reach the required number of characters.
9. If you fail to find a match, start again with the current first bigram.
10. Write out the output text.

There are sample generated texts in the accompanying dataset.

## 5. List of datasets and files

The following files are included in the related .zip file.

<i>File</i>	<i>Description</i>
cervantes-10k.txt	Cervantes' Clever Writer 0
cervantes-20k.txt	Cervantes' Clever Writer 1
cervantes-30k.txt	Cervantes' Clever Writer 2
cervantes-40k.txt	Cervantes' Clever Writer 3
cervantes-50k.txt	Cervantes' Clever Writer 4
cervantes-60k.txt	Cervantes' Clever Writer 5
cervantes-100k.txt	Cervantes' Clever Writer 6
cervantes-1mb.txt	Cervantes' Clever Writer 7
cervantes-10k.freq.txt	Frequency analysis of Cervantes' Clever Writer 0
cervantes-20k.freq.txt	Frequency analysis of Cervantes' Clever Writer 1
cervantes-30k.freq.txt	Frequency analysis of Cervantes' Clever Writer 2
cervantes-40k.freq.txt	Frequency analysis of Cervantes' Clever Writer 3
cervantes-50k.freq.txt	Frequency analysis of Cervantes' Clever Writer 4
cervantes-60k.freq.txt	Frequency analysis of Cervantes' Clever Writer 5
cervantes-100k.freq.txt	Frequency analysis of Cervantes' Clever Writer 6
cervantes-1mb.freq.txt	Frequency analysis of Cervantes' Clever Writer 7
clean-spanish-book-extracts.txt	Extracts for Gutenberg
spanish_leipzig-clean.txt	Leipzig corpus, cleaned.
make-cervantes-monkey.php	Sample code to generate texts
spanish-letters-space-unicase-leipzig-gutenberg.ods	Unicase letter frequencies
spanish-bigrams-unicase-letters-leipzig-gutenberg.ods	Unicase bigram counts
spanish-words-200.csv	Most common words

## 6. Acknowledgements

Thanks to (alphabetically) Arno Klein, Miguel Guzmán, and Nick Gutiérrez for their inputs and assistance. Also to the team behind the Libertinus fonts. [13]

## 7. Bibliography

- [1] A. Klein, *Engram-es Spanish keyboard layout*. 2021. Accessed: Sep. 04, 2021. [Online]. Available: <https://github.com/binarybottle/engram-es>
- [2] A. Klein, ‘Engram: A Systematic Approach to Optimize Keyboard Layouts for Touch Typing, With Example for the English Language’, Mar. 2021, doi: 10/gjj5kh.
- [3] ‘NickG13 - Overview’, GitHub. <https://github.com/NickG13> (accessed Sep. 04, 2021).
- [4] ‘Lobo-Feroz - Overview’, GitHub. <https://github.com/Lobo-Feroz> (accessed Sep. 04, 2021).
- [5] S. Trost, ‘Alphabet and Character Frequency: Spanish (Español)’. <https://www.sttmedia.com/characterfrequency-spanish> (accessed Sep. 04, 2021).
- [6] ‘N-grams: based on one billion word COCA corpus’. <https://www.ngrams.info/spanish.asp> (accessed Sep. 04, 2021).
- [7] I. Douglas, ‘Keyboard Layout Analysis: Creating the Corpus, Bigram Chains, and Shakespeare’s Monkeys’, Zenodo, Mar. 2021. doi: 10.5281/zenodo.4644104.
- [8] ‘Download Corpora Spanish’. <https://wortschatz.uni-leipzig.de/en/download/Spanish> (accessed Sep. 04, 2021).
- [9] D. Goldhahn, T. Eckart, and U. Quasthoff, ‘Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages’, p. 7.
- [10] ‘Project Gutenberg’, Project Gutenberg. <https://www.gutenberg.org/> (accessed Sep. 04, 2021).
- [11] ‘List of spanish words | listapalabras’. <https://www.listapalabras.com/en/> (accessed Sep. 05, 2021).
- [12] M. Gerlach and F. Font-Clos, ‘A standardized Project Gutenberg corpus for statistical analysis of natural language and quantitative linguistics’, *arXiv:1812.08092 [physics]*, Dec. 2018, Accessed: Mar. 27, 2021. [Online]. Available: <http://arxiv.org/abs/1812.08092>
- [13] C. MacLennan, *Llibertinus font*. 2020. Accessed: Jan. 01, 2021. [Online]. Available: <https://github.com/alerque/libertinus>