Generación de Diceware (DW) en Español

2003-05-10

Manuel Palao, CISM, CISA

DW-Español-1.txt
y
DW-Español-2.txt

DW-Español-1.txt y DW-Español-2.txt son 2 listas de 7,776 palabras principalmente españolas [1] elaboradas para los usuarios hispanoparlantes del método Diceware (http://world.std.com/~reinhold/diceware.html) de generación de frases de paso seguras. [Primero creé DW-Español-2.txt y luego DW-Español-1.txt].

Ambas listas son plenamente compatibles con el método Diceware y son criptográficamente equivalentes. La diferencia fundamental entre ambas es que la segunda contiene caracteres diacríticos (tildes de acentos) lo que en mi opinión la hace más mnemotécnica (para quienes respetan la ortografía), y da, en principio [2], lugar a una longitud media de palabra inferior; el inconveniente es que se limita su compatibilidad entre plataformas.

Elaboración de la lista DW-Español-2.txt:

1. Muy similar a la propuesta en Diceware FAQs.

2. Comencé bajándome una lista de 58.485 palabras españolas de http://sinetgy.org/pipermail/dino/2002-February/000162.html. Había localizado en Google otras listas, pero la mayoría de ellas carecían de tildes de acentos, que a mí me parecían importantes por corrección ortográfica y por mnemotecnia.

3. Esa lista constaba fundamentalmente de palabras en minúsculas, aunque incluía algunas que comenzaban por mayúscula (patronímicos y toponímicos).

[1] Ver punto 10.

[2] Finalmente la Lista DW-Español-1.txt acabó teniendo una longitud media de palabra inferior a la Lista DW-Español-2.txt, gracias a los 'mini-chunks' a que me refiero más adelante.

4. Usé en Excel una función para contar los caracteres, y ordené por longitud creciente.

5. La longitud de las palabras iba de 1 a 21 letras [con 21 sólo 1: 'electroencefalografía'].

6. Había 5.222 palabras de hasta (incl.) 5 letras (las salvé como Lista A) y 11.363 de hasta 6 letras (salvé las de 6 letras como Lista B).

7. Ojeando la Lista A excluí –no de modo sistemático [3]— las palabras que consideré obviamente potencialmente ofensivas y otras aún no aprobadas por el Diccionario 2001 de la RAE (http://www.rae.es/), o que consideré demasiado dialectales o locales. Dejé, sin embargo un cierto número de palabras locales o de oscuro significado (¡local y oscuro son relativos!) como'salep', 'satis', 'segrí'; 'sariá' (Argentina) y 'tucán' (Brasil). La idea es que entre 7.776 palabras siempre habría algunas desconocidas para el usuario que para memorizarlas debería –en todo caso— consultarlas en el diccionario, para conservar una memoria semántica, o memorizar su grafismo.

8. Añadí a la Lista A ciertas palabras que opiné eran transformaciones legítimas (según la Gramática de la RAE), como plurales, o poner o quitar la inicial mayúscula a ciertas palabras : Adán, adán; sol, Sol.

9. De la Lista B seleccioné arbitrariamente unas 1.040 que me parecieron populares, y obtuve la lista C.

10. De modo independiente creé, partiendo de cero, una Lista D con 1.837 "palabras", muy similares a las «Diceware's 992 "word" file dicewarekit.txt», con algunas diferencias significativas en el enfoque (con lo que conseguí el doble de palabras). Incluí:

    10.1. números.
    10.2. ordinales (3º) en masculino y femenino (3ª).
    10.3. cadenas cortas de todos los caracteres especiales del teclado de PC estándar español [4].
    10.4. usos 'legítimos' de los símbolos diacríticos (*acute accent, á; dieresis*, etc.).
    10.5. Nombres (con inicial mayúscula) de personas, lagos, cordilleras, países, ciudades, mares, ríos, etc. [NUEVO].

---

[3] Quiero señalar que no tengo ninguna preparación formal en Lingüística, por lo que muchas de mis decisiones han podido ser poco rigurosas, aunque pienso que suficientemente correctas para el objetivo: obtener 7.776 palabras 'buenas'. Hace unos años incurrí en un esfuerzo similar (mucho más reducido) para preparar una lista de palabras breves en español para un sistema de protección de distribución de software con InterLok de PACE (paceap.com), del que adquirí cierta experiencia.

[4] *MS-DOS Codepage 850 (Multilingual Latin 1) characters:*
199 LATIN CAPITAL LETTER C WITH CEDILLA
231 LATIN SMALL LETTER C WITH CEDILLA,
252 LATIN SMALL LETTER U WITH DIERESIS

10.6. Nombres (con inicial mayúscula) de famosos actores, escritores, pintores, héroes mitológicos, planetas, personajes de novela o película (Nemo, Tarzán) [5]. [NUEVO].
10.7. símbolos químicos comunes como H2O. [NUEVO].
10.8. marcas como KODAK, IBM o BMW. [NUEVO].
10.9. palabras inglesas popularizadas ('boy', 'bus', ..., 'Yankee'). [NUEVO].

11. Combiné entonces las Listas A, C y D, comprobé que no había palabras repetidas [6], pero retuve las que se diferenciaban en la inicial mayúscula o no y en las tildes de acente, y tenían sentido distinto. Ej.: 'dólar' ($) y'dolar' (pulimentar madera o piedra).

12. En todo el proceso intenté combinar dos objetivos: brevedad de la palabra media y facilidad de recordarla (esto último es obviamente subjetivo). Perseguí on equilibrio entre el enfoque tradicional de Diceware (ni mayúsculas ni tildes) y la brevedad de la palabra y su correcta grafía en español.

13. Obtuve así 7,776 palabras, que creo que son breves y no repetidas. Añadí entonces la columna de 5 números tomada de Diceware.

14. El análisis de los resultados es el siguiente:

# letras	# palabras
1	75
2	300
3	1.432
4	1.449
5	3.429
6	1.091
TOTAL	7.776

La longitud media de la palabra es 4,4, próxima al 4,2, del Diceware original, un buen resultado, si se considera que la verbosidad (longitud de palabras, longitud de frases) del español es considerablemente superior a la del inglés.

Para una frase de paso de 5 ó 6 palabras, la longitud media (añadiendo 4 ó 5 espacios entre palabras) es de 22,0 y 31,4 caracteres, respectivamente.

15. Caracteres

He usado MS-DOS Codepage 850 (Multilingual Latin 1).

Esto añade a la lista Diceware, como palabras reales, 15 nuevos caracteres al conjunto estándar de 26 letras en inglés ,y 26 mayúsculas potenciales más.

[5] Dejé fuera a Tolkien, porque él sólo hubiera copado la lista ;)

[6] Por cierto, en Excel, ' = ' no sirve para esto porque ignora mayúsculas o minúsculas. Hay que usar la función IGUAL.

Code	ISO/IEC 10646-1:2000 Character Name
193	LATIN CAPITAL LETTER A WITH ACUTE
199	LATIN CAPITAL LETTER C WITH CEDILLA
201	LATIN CAPITAL LETTER E WITH ACUTE
205	LATIN CAPITAL LETTER I WITH ACUTE
209	LATIN CAPITAL LETTER N WITH TILDE
211	LATIN CAPITAL LETTER O WITH ACUTE
218	LATIN CAPITAL LETTER U WITH ACUTE
225	LATIN SMALL LETTER A WITH ACUTE
231	LATIN SMALL LETTER C WITH CEDILLA
233	LATIN SMALL LETTER E WITH ACUTE
237	LATIN SMALL LETTER I WITH ACUTE
241	LATIN SMALL LETTER N WITH TILDE
243	LATIN SMALL LETTER O WITH ACUTE
250	LATIN SMALL LETTER U WITH ACUTE
252	LATIN SMALL LETTER U WITH DIERESIS

Así las palabras de mi lista contienen hasta 26 + 26 + 15 = 67 caracteres distintos. Sin embargo, la frecuencia de los 41 caracteres añadidos es muy baja.

16. Entropía y fuerza de la frase de paso

Si entiendo correctamente la matemática subyacente:

16.1. La fuerza criptográfica de la frase de paso depende de su método de generación y de la información que el atacante tenga del mismo o las hipótesis correctas que haga.

16.2. La entropía de mi versión de Lista Diceware en Español es de 12,9 bits por palabra [$\log_2(7776)$], de modo que una frase de paso de 5 palabras tendría una entropía mínima de 64,5 bits, que sería la misma independientemente de la Lista Diceware utilizada. Si el atacante supone que la frase de paso pertenece a Lista Diceware en Español, su entropía teórica (ver 16.3) la reduce a 64,5 bits.

En este caso, el esfuerzo para un 'ataque de diccionario' comenzaría explorando una frase de paso de 1 palabra, luego 2 palabras ... hasta 5. Y al 50% de la exploración de 5 palabras acertaría, como media:

$7.776 + 7.776^2 + 7.776^3 + 7.776^4 + 0{,}5 \times 7.776^5 = 1{,}42E{+}19$ comprobaciones de palabras. Esto es 4,51E+05 (451.000) MIPS-año [7], suponiendo una instrucción por comprobación [8].

16.3. La "entropía teórica" (para un 'ataque de fuerza bruta') es la de 22 caracteres (para 5 palabras de un alfabeto de 67), esto es $\log_2(67^{22}) =$

[7] 1 MIPS = 3.15E+13 instrucciones por segundo; 1 MIPS-año = 1 MIPS durante 1 año.

[8] RSA-140 fué factorizado con unos recursos estimados de 2,000 MIPS-año. http://www.counterpane.com/crypto-gram-9903.html#RSA140

133,45 bits. Este número es una cota superior facilona, ya que la frecuencia real de los 67 caracteres varía significativamente, y los 41 añadidos están infrarepresentados.

16.4. Un 'ataque de diccionario' requeriría, para mi versión, el uso de 2 diccionarios (español e inglés básico [9]). Supongamos 70.000 + 1.000 = 71.000 palabras.

El ataque (como con cualquier Lista Diceware) comenzaría explorando una frase de paso de 1 palabra, luego 2 palabras ... hasta 5. Y al 50% de la exploración de 5 palabras acertaría, como media. (Estoy soslayando el hecho de que en la frase de paso hay espacios entre palabras, y el hecho significativo de que la Lista contiene más de 1.400 (18%) "no palabras" (no del lenguaje natural normalizado: marcas y cadenas de caracteres especiales).

El esfuerzo necesario, ignorando el trabajo preparatorio de la Nota 6, sería:

$71.000 + 71.000^2 + 71.000^3 + 71.000^4 + 0{,}5 \times 71.000^5 = 9{,}02E+23$ comprobaciones de palabras. Esto es, 2,86E+10 (28.600 millones) MIPS-año. ¡ Suponiendo una instrucción por comprobación !

* * *

Generación de DW-Español-1.txt (la Segunda Lista generada):

Seguí la recomendación de Arnold Reinhold [10] –el 'padre' de Diceware— de prescindir de palabras con caracteres que no fueran ASCII 0:127. ¡Había 1.370, demasiadas –en mi opinión— para substituirlas por palabras de 6 letras, así que decidí substituirlas por palabras 'no naturales' más cortas. Generé el patrón siguiente: letra minúscula, seguida por 2 dígitos, el primero 1:9; el segundo 0:9. Descarté ciertas letras 'equívocas', como i. l. o. etc. Creo que ese subconjunto contiene 'mini-chunks' muy fáciles de recordar.

El resultado es el fichero DW-Español-1.

La mejora en longitud de palabra ha sido significativa:

[9] Determinar la necesidad de 2 diccionarios probablemente implicaría previamente dos conjeturas correctas y un ataque de diccionario: i) la conjetura (no demasiado difícil de hacer acertadamente, si se conoce por ejemplo en qué idioma me comunico habitualmente)de que mi Lista está en español); ii) un ataque de diccionario español, fallido; y iii) la conjetura de que parte de mi Lista está en inglés (no demasiado difícil de hacer y que concluiría que se trata de una lista básica corta). .

[10] Quiero agradecer desde aquí a Arnold Reinhold sus autorizaciones, estímulos y críticas.

Longitud	Número	%
1	56	0,72
2	251	3,23
3	2581	33,19
4	1258	16,18
5	2804	36,06
6	826	10,62
Total	7776	100,00

Media 3,26
5 palabras + espacios: 20,28

* * *