Reconocimiento de voz de Windows Vista: ¿mejor, igual o peor que Dragon Naturally Speaking?

Lorenzo Serrahima*

Resumen: La nueva versión de Windows, Windows Vista, incorpora una utilidad llamada reconocimiento de voz (RVV). El presente artículo describe una prueba realizada con la intención de responder a la pregunta de si es mejor o peor que el programa Dragon Naturally Speaking (DNS). Se ha diseñado una prueba comparativa entre ambos, procurando evitar al máximo cualquier sesgo que pudiera restar fiabilidad a la comparación. De los resultados de la prueba se deduce que con ambos programas se puede alcanzar una precisión superior al 95 %, sin que se hayan detectado diferencias significativas de calidad entre uno y otro.

Palabras clave: reconocimiento de voz, programas, prueba comparativa

Windows Vista's Voice-recognition Software: better, same or worst than Dragon Naturally Speaking?

Abstract: TThe new version of Windows, Windows Vista, includes a tool called Vista Voice Recognition (VVR). This article describes a test conducted to find out whether it is better or worse than Dragon Naturally Speaking (DNS) software. To this end, a benchmark was designed in an attempt to minimize any possible bias which could reduce its reliability. The results show that the precision of both programs is more than 95%, and no findings of any significant differences in quality.

Key words: voice recognition, software, benchmark.
Panace@ 2009; 10 (29): 76-79

Introducción

La nueva versión de Windows, Windows Vista, incorpora una utilidad llamada reconocimiento de voz (RVV). De inmediato surge la pregunta de si le será útil al traductor profesional como programa de dictado o no.

Hasta ahora, en el mercado había pocos programas de reconocimiento de voz que además tuvieran utilidad para los traductores. El más conocido y más extendido de ellos es Dragon Naturally Speaking (DNS). Dado que es bastante conocido entre la comunidad de traductores, se ha utilizado como referencia con la que comparar la eficacia de la nueva herramienta RVV de Windows. Se describe una prueba comparativa entre ambos, y se procura evitar al máximo cualquier sesgo que pudiera restar fiabilidad a la comparación.

Material y métodos

Se ha utilizado un ordenador de sobremesa Fujitsu Siemens, CPU Pentium IV 2.40 GHz con 1,25 GB de RAM, dotado de dos monitores y micrófono NGS con auriculares. Contenía el sistema operativo Microsoft Windows XP y programas Adobe Reader 7.0 (para mostrar el texto original), Microsoft Word 2002 (con el que se escribía la traducción) y Dragon Naturally Speaking 9.0 (para el dictado).

Por otro lado, se ha utilizado un ordenador portátil Acer Aspire 9303WSMi, CPU AMD Turion 64 x2 TL-52 1,6 Ghz con 2 GB de RAM con pantalla de 17" y micrófono con auriculares sin marca. Tiene el sistema operativo Microsoft Windows Vista (con la función reconocimiento de voz activada para el dictado) y el programa Microsoft Word 2002 (con el que se escribía la traducción).

Los dos micrófonos habían sido comprados en distintas ocasiones en sendas tiendas de informática; ambos eran de la gama baja de precios y son los que habitualmente utiliza el autor en su trabajo cotidiano.

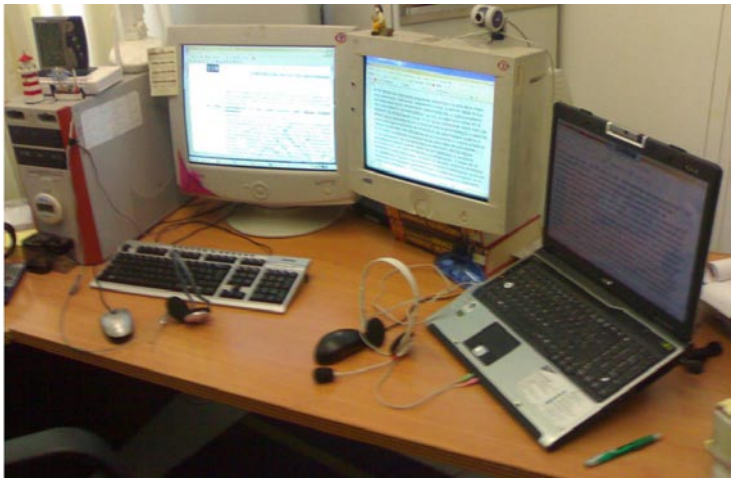
La prueba consistió en dictar un texto de unas 3300 palabras simultáneamente con ambos programas. Se colocaron los dos ordenadores sobre una mesa y se conectó un micrófono con auriculares a cada uno de ellos. El autor se colocó ambos auriculares simultáneamente, de manera que los dos micrófonos quedasen a la misma distancia de la boca, tal como se aprecia en la fotografía que aparece en la página siguiente. En el ordenador de sobremesa se mostraba el texto original a través de uno de los monitores y el programa de tratamiento de texto en el otro. En la pantalla del ordenador portátil se mostraba el programa de tratamiento de texto.

El texto era un caso real de traducción de un informe médico del alemán al español, para que las condiciones de la prueba fuesen lo más parecidas posible a las condiciones reales de uso.

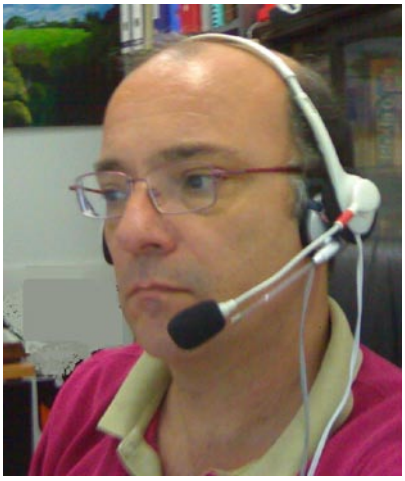
El programa de referencia era el DNS que el autor lleva seis años utilizando habitualmente, primero la versión 4.0, después la versión XP y actualmente la versión 9.0, que funciona con el sistema operativo Windows XP. A lo largo de los años ha ido incorporando abundante vocabulario médico al vocabulario común inicial que traía de fábrica.

El dictado se dividió en tres partes. En ninguna de ellas se introdujo ningún cambio en el programa DNS, pero sí

* Veterinario, traductor autónomo, Barcelona (España). Dirección para correspondencia: serrahima@gmail.com.



Los dos ordenadores



Detalle de los dos micrófonos

den el RVV. La primera parte fue dictada con el RVV sin realizar ningún tipo de entrenamiento previo. En principio, esto debería ir en detrimento de la precisión del programa RVV, que cabría esperar que fuese menor que la alcanzada con el DNS. Al acabar, se hizo el entrenamiento que supuestamente debía mejorar significativamente la precisión del programa. Este punto es importante, porque aquí se iba a comprobar si la precisión aumentaba significativamente y si alcanzaba o incluso superaba la obtenida con DNS. A continuación se dictó la segunda parte. Finalmente se intercambió la conexión de los micrófonos, para descartar que pudieran influir en la calidad del resultado, y se dictó la tercera parte. Las tres pruebas se hicieron seguidas y ocuparon aproximadamente tres horas. En ningún caso se hizo corrección alguna durante el dictado, pero, una vez acabado, se borraron o modificaron todos los apellidos de las personas y los nombres de las instituciones y ciudades que se citan en el informe para garantizar la necesaria confidencialidad de la información. Los textos dictados se guardaron sin revisar, para su posterior comparación.

Cada uno de los textos dictados se dividió en las tres partes de que constaba la prueba. Se hizo una segmentación por frases y a continuación cada parte se transformó en una tabla. De esta manera se facilitaba mucho la revisión de los textos y el marcado de los errores. Se hizo una primera revisión en la que se marcaron los errores detectados y, a continuación, se definieron de la siguiente forma:

A. Plurales, acentos, mayúsculas. Incluye palabras escritas correctamente, prácticamente iguales a las que se han dictado, pero con alguna variación, como escribirlas indebidamente con mayúscula o minúscula (p. ej., *Hermano*), colocación de acentos distintos a los que se han dictado (p. ej., *él* en lugar de *el*), escritura en plural o singular (p. ej., *pruebas* cuando se ha dictado *prueba*), tiempos verbales distintos a los dictados (p. ej., *terminar* en lugar de *termina*).
B. Incorrectas comunes. Palabras pertenecientes al lenguaje común distintas a las dictadas, correcta o incorrectamente escritas (p. ej., *culebra* cuando se ha dictado *prueba*).
C. Incorrectas técnicas. Palabras pertenecientes al lenguaje técnico distintas a las dictadas, correcta o incorrectamente escritas (p. ej., *euros psicológica* cuando se ha dictado *neuropsicológica*). En este grupo de palabras es fácil reducir el número de errores, basta con incorporar las palabras desconocidas a la memoria del programa. Sin embargo, por la propia dinámica de la prueba no se hizo, de ahí que muchos de los errores de esta categoría afectasen casi siempre a las mismas palabras *(neuropsicología, parafasia…).*
D. Olvidadas o añadidas. Palabras que se han dictado y no se han escrito, o palabras cortas añadidas por el programa sin que hayan sido dictadas (p. ej., *respecto a a* cuando se ha dictado *respecto a*).
E. Siglas y números. Esta categoría abarca todo tipo de siglas que se hayan podido dictar (p. ej., TAV) y los números de cualquier clase. El hecho de que un número se escribiera con cifras o con letras no se ha considerado un error si el número era el dictado. Es decir, tras dictar el 2 se consideraba un acierto tanto el 2 como la palabra *dos.*
F. Signos ortográficos. Esta categoría abarca los signos de puntuación, paréntesis, guiones y demás signos utilizados en la ortografía habitual. El hecho de que no se escribiera el signo, sino la palabra que lo define, no se consideraba error. Es decir, tras dictar «guión», se consideraba un acierto tanto el «-» como la palabra *guión.* Sin embargo, si aparecía otra palabra (p. ej., *billón*), se consideraba un error. Es importante destacar que las órdenes utilizadas en el dictado fueron siempre las del programa DNS. Es decir, para cambiar de línea la orden usada fue «nueva línea», y para poner comillas las órdenes fueron «abrir comillas» y «cerrar comillas». Dado que no todas las órdenes son idénticas para ambos programas, era de esperar que con RVV apareciesen más errores en la categoría de signos ortográficos que con DNS.

Después de esto, se hizo una segunda revisión de los mismos textos marcando con detalle todos los errores y haciendo su correspondiente recuento. Lo ideal hubiese sido poder hacer un análisis estadístico para comprobar si las diferencias entre un programa y otro son estadísticamente significativas. Pero para que este tipo de análisis tenga sentido es imprescindible que la única condición distinta sean los programas que se están probando, cosa que no se cumplió en esta prueba: los dos ordenadores eran de marca y configuración diferentes, con un sistema operativo diferente. En caso de aparecer diferencias estadísticamente significativas quedaría la duda de si eran debidas al programa o a algún otro factor. De ahí que no se haya hecho ningún análisis estadístico. El único factor cuya influencia pudo reducirse fue precisamente el de los micrófonos, pues en la parte III de la prueba se intercambiaron.

Resultados

En todos los casos se considera el valor obtenido con el programa DNS como el valor de referencia, es decir, el 100 %. El texto final tuvo una extensión muy similar con ambos programas (DNS: 3311 palabras; RVV: 3356 palabras [101,36 %]). En ninguna de las tres partes de la prueba la diferencia del número de palabras alcanzó el 3 %. Los distintos errores cometidos por cada uno de los programas explican que, a pesar de ser el mismo texto para ambos, el número de palabras finales sea distinto. En el cuadro 1 se detalla el número de errores cometidos por cada programa, clasificados por partes de la prueba y por categorías.

Parte I

En esta primera parte de la prueba todavía no se había hecho el entrenamiento con el programa RVV, y se empezó a dictar tal como venía instalado en el ordenador. Por eso cabía esperar un número de errores relativamente elevado y, de hecho, así fue: el texto dictado con RVV tenía casi un 50 % más de errores que el dictado con DNS (75,77 ‰ frente a 52,95 ‰). Las diferencias más importantes se observan en el apartado B, palabras comunes, aunque sorprendentemente RVV cometió menos de la mitad de errores que DNS. También sorprende que el número de errores en el apartado C, palabras técnicas, sea igual para ambos programas. Supuestamente, DNS debería ser capaz de identificarlas mejor, dado que su base de datos se ha ido ampliando con palabras

Cuadro 1: Errores cometidos por cada programa en cada prueba

PALABRAS	DNS	RVV	(‰) DNS	(‰) RVV
PARTE I	1001	1023 (+2,20%)		
A - Plurales, acentos, mayúsculas	2	6	2,00	5,87
B - Incorrectas comunes	17	7	16,98	6,84
C - Incorrectas técnicas	26	25	25,97	24,44
D - Olvidadas o añadidas	2	9	2,00	8,80
E - Siglas y números	3	9	3,00	8,80
F - Signos ortográficos	3	21	3,00	20,53
Errores totales	*53*	*77*	*52,95*	*75,27*
PARTE II	1030	1016 (-1,36%)		
A - Plurales, acentos, mayúsculas	4	6	3,88	5,91
B - Incorrectas comunes	5	6	4,85	5,91
C - Incorrectas técnicas	9	10	8,74	9,84
D - Olvidadas o añadidas	4	2	3,88	1,97
E - Siglas y números	1	1	0,97	0,98
F - Signos ortográficos	8	1	7,77	0,98
Errores totales	*31*	*26*	*30,10*	*25,59*
PARTE III	1280	1317 (+2,89%)		
A - Plurales, acentos, mayúsculas	13	9	10,16	6,83
B - Incorrectas comunes	14	9	10,94	6,83
C - Incorrectas técnicas	18	19	14,06	14,43
D - Olvidadas o añadidas	7	9	5,47	6,83
E - Siglas y números	2	5	1,56	3,80
F - Signos ortográficos	3	12	2,34	9,11
Errores totales	*57*	*63*	*44,53*	*47,84*

técnicas nuevas durante varios años. Aunque el autor no tiene una explicación definitiva para este fenómeno, ha percibido un funcionamiento distinto en los dos programas que podría explicarlo. DNS identifica los sonidos que oye del usuario, los divide en palabras, compara cada una de las palabras identificadas con lista de palabras del propio programa, escoge la que más se parece al sonido identificado y la coloca en su lugar en el texto. Es decir, para que DNS pueda escribir correctamente una palabra, previamente debe tenerla escrita en su base de datos. Sin embargo, RVV parece funcionar de otra forma. Aparentemente, RVV identifica los sonidos que oye del usuario, los convierte en sílabas y después agrupa las sílabas por palabras. Este proceso parece ser más eficaz en el caso de las palabras desconocidas para el programa, sobre todo cuanto más largas son. Eso le permite acertar correctamente un buen número de palabras técnicas desconocidas para el programa.

En el apartado E, palabras olvidadas o añadidas, es lógico que el número de errores de RVV sea mayor. Al no haber hecho el correspondiente entrenamiento, el programa no identifica correctamente todos los matices de la voz de quien dicta y se pierden algunas palabras. Lo mismo sucede en el momento de dictar siglas o números, si bien aquí cabe citar un error de RVV: en ocasiones escribe los números romanos (aleatoriamente, sin que se hayan dictado así) con letras minúsculas. Podemos decir que este es un error estructural, pues aunque los anglosajones usen las minúsculas para los números romanos, la RAE, en su *Diccionario panhispánico de dudas*, dice claramente que los números romanos deben escribirse siempre con mayúsculas.[1] Por lo tanto, cabe exigir al fabricante que, en la versión para dictar en español, el programa use exclusivamente las letras mayúsculas para ellos. No se han detectado más errores estructurales, pero la prueba realizada no es lo bastante grande como para considerar que se han probado todas las posibilidades del programa.

Como cabía esperar, la cifra de errores con los signos ortográficos es mucho más elevada con el programa RVV.

Parte II

Antes de dictar esta segunda parte, se hizo el entrenamiento del programa RVV, con lo que cabría esperar que mejorase la calidad del texto dictado, como así fue. Es probable que el texto en su conjunto fuese más «fácil de entender» para un programa de dictado, puesto que la proporción de errores cometidos por DNS fue menor en esta segunda parte (52,95 ‰ en la primera, 30,10 ‰ en la segunda). Los resultados obtenidos con el programa RVV fueron muy similares en su conjunto, aunque destaca y sorprende la diferencia en los errores ortográficos, pues fueron mucho más frecuentes con el programa DNS. Hubiese sido muy interesante poder «trasladar» la experiencia adquirida con DNS al programa RVV, mediante una función de incorporación de listas de palabras similar a la que ofrece DNS. Sin embargo, el programa RVV solamente permite incorporar palabras nuevas a su diccionario de una en una, lo que hace inviable incorporar las más de 3000 palabras nuevas con las que el autor ha ampliado el diccionario de DNS.

Parte III

En esta tercera parte se intentó comprobar si el micrófono tenía alguna influencia significativa sobre la calidad de los resultados o no. Para ello se intercambiaron los micrófonos entre ambos ordenadores. El hecho de que esta parte sea un 30 % más larga (unas 1300 palabras) es puramente circunstancial. Por otro lado, parece que el texto era algo más «difícil de entender» para un programa de dictado, puesto que la proporción de errores cometidos por ambos programas aumentó. DNS pasó de un 30,10 ‰ en la parte II a un 44,53 ‰ en la parte III, y RVV pasó de un 25,59 ‰ en la parte II a un 47,84 ‰ en la parte III. Sin embargo, ambos programas alcanzaron un índice de precisión muy similar. La única diferencia apreciable radica en los errores ortográficos, esta vez mucho más frecuentes con el programa RVV. Aunque se hace difícil atribuir el cambio en los errores ortográficos al cambio de micrófono, en realidad es la única variable modificada entre una prueba y otra. La única hipótesis que quizá explicaría este fenómeno sería que uno de los dos micrófonos (de distinta marca y forma, como se observa en la fotografía) fuese más sensible a los ruidos ambientales que el otro. Ambos estaban montados uno sobre otro, y por lo tanto es probable que al mover la cabeza o cambiar ligeramente de postura se produjeran ruidos de roce entre ambos. Quizá uno de los dos micrófonos los detectaba y estos ruidos interferían con mayor intensidad en las órdenes dadas al programa que en las palabras.

Conclusiones

Una vez realizada la prueba y valorados todos los resultados, se comprueba que, en las condiciones de trabajo habituales, los dos programas son capaces de alcanzar una precisión muy similar en el texto dictado. Con muy poco entrenamiento, ambos son capaces de escribir correctamente más del 95 % de las palabras dictadas. El programa RVV parece tener una precisión ligeramente mayor con las palabras comunes, pero esta diferencia es poco marcada, y por lo tanto resulta difícil valorar si es significativa o casual. En cualquier caso, en esta prueba el RVV ha demostrado ser un programa de dictado de calidad similar al DNS; es decir, la función de reconocimiento de voz de Windows Vista ofrece resultados más que satisfactorios para el trabajo diario de dictado de un traductor, lo que hace innecesario añadir otro programa de dictado.

Nota del autor: Todo el texto de este artículo ha sido dictado con DNS y revisado posteriormente con el teclado.

Notas

[1] <http://buscon.rae.es/dpdI/>. Entrada «números».