El Software libre y la lingüística

Maria Francisca Ribeiro de Araujo Santo Orcero

FCLAR/UNESP (Brasil)

fran@orcero.org

David Santo Orcero

Consultor de soluciones con software libre

irbis@orcero.org

La sociolingüística es un área en la que la informática aún no ha penetrado completamente. Los autores de este trabajo hemos intentado informatizar una investigación sociolingüistica completa usando software libre en todos los lugares donde esto ha sido posible, incluso implementando software en alguno de los pasos. En este trabajo estudiaremos las ventajas de la informatización con software libre de la sociolingüsitica, qué software está disponible, cual ha sido nuestra experiencia, y aquellos puntos donde todavía no existe reemplazo al software propietario.

1. Introducción al problema de la informática y

la sociolingüística

La sociolingüística se encuentra con dos problemas serios en la investigación de campo, que son la grabación y el almacenamiento de datos del audio. Hasta ahora, la grabación y el almacenamiento de datos de investigaciones de campo en lingüística se ha realizado por medio de las cintas cassettes. Esto hace al procedimiento de transcripción fonética extremadamente complejo y engorroso, debido al ruido propio de las cintas, a la pérdida de calidad de las grabaciones por su uso, con la pérdida de datos invaluables para la ciencia de hablas, de acentos y hasta de lenguas que han desaparecido o están en vias de desaparición, y las cintas con las conversaciones con los hablantes se estan degradando, perdiendo toda la información.

La propia investigación y transcripción fonética de las cintas es destructiva: el movimiento de ir y venir con la cinta cassette, muchas veces, causa la ruptura de la cinta y la pérdida irrecuperable de los datos grabados. Hacer copias múltiples de las cintas cassettes presenta disminuciones de la calidad de la cinta original, además de que la copia es siempre de peor calidad que el original; además de esto, las cintas son vulnerables al moho con el tiempo cuando no son bien conservadas. Dependiendo del alcance de la investigación, el número de cintas puede llegar a cantidades realmente enormes y la gestión de estos grandes volúmenes de datos de audio se complica mucho.

Por último, a pesar de que automatizáramos el proceso de recogida de datos, el procesamiento de los datos es aún engorroso y propenso a fallos. El único programa existente que estudia las correlaciones entre datos lingüisticos, el VARBRUL, es un programa de MS-DOS de código cerrado, lento y muy poco amigable para el usuario.

Este trabajo también corresponde al aspecto informático de una investigación realizada sobre un dialecto hablado en Caxias, Brasil, una pequeña ciudad de 40000 habitantes, la mayor parte de ellos ancianos, por un impresionante flujo migratorio de los jóvenes a ciudades que presentan posibilidad de empleo, que ha hecho que la población de la ciudad caiga a su tercera parte en quince años. Este trabajo de investigación ha sido realizado en su integridad con herramientas libres, para analizar la posibilidad de informatizar todo el proceso de colecta y gestión de datos, así como publicación de los resultados usando solo software libre.

Este trabajo ha sido financiado parcialmente por la FAPESP, organización de la que MFRASO es becaria de investigación.

2. Descripción del problema de las cintas

La cinta ha sido hasta el momento un elemento indispensable en las grabaciones de

datos sociolingüísticos. No queremos negar su gran importancia en el pasado, pero tampoco queremos negar algunos problemas inherentes a su uso, entre los que los más comunes son:

- Las cintas se estropean fácilmente con el movimiento continuado de avanzar y retroceder.

  "¡La cinta se rompió exactamente en el lugar donde yo necesitaba oír la grabación una vez más!" "¿Y ahora? ¡Los hablantes hace tiempo que murieron!" "Perdí el trabajo de un año! Yo intenté encolar los puntos de la cinta con un cinta adhesiva, pero no se quedó bien". Ésas son algunas frases de desesperación por perder los datos de una investigación por rotura de la cinta. Obtener nuevos datos no es una tarea fácil, y en el caso de comunidades de difícil acceso, en vías de extinción o extintas es imposible, y esa parte de la cultura de la humanidad se habrá perdido para siempre. Ir al campo presupone disponer de tiempo, paciencia y habilidad de trabajar con una comunidad de hablantes (Labov 1994). Muchas veces los hablantes no aceptan a ser entrevistados temerosos de represalias políticas, lo que hace los datos difíciles de conseguir aunque la comunidad que posea esa variante siga viva.

- Con el tiempo, las cintas van perdiendo en calidad, aunque no se usen.

  Aunque se tomen los cuidados necesarios en la conservación de las cintas magnéticas, el tiempo acaba por destruir la calidad de las cintas y esto es inevitable.

- Las cintas son sensibles a la humedad, al calor y los campos magnéticos, aunque sean campos pequeños.

  El moho es el principal enemigo de las cintas magnéticas que se quedan guardadas por mucho tiempo, llegando incluso a destruirlas. Para resolver ese tipo del problema, es importante que un especialista realice una limpieza periódica de la superficie de la cinta. Aun así, una limpieza cuidadosa es económicamente inviable, por la gran cantidad de metros de cinta involucrados. Por ello, los datos terminan perdiéndose dentro de los laboratorios.

- Las cintas ocupan mucho espacio físico.

  Para grabar un hablante, se usa una cinta de 60 minutos, por lo menos. Multiplicando esos minutos por 12, para construir la muestra de investigación más simple posible con representatividad, tendremos el equivalente de 720 horas de grabaciones que ocuparán 12 cintas, por lo menos. Si la muestra crece, como son las muestras

dialetológicas (cf. Ferreira &Cardoso 1994), esa equivalencia se triplica y los perjuicios serán, entonces, la falta del espacio en los laboratorios, la conservación de las cintas (comentado en (c)) y su distribución.

- La copia es siempre peor que el original.

  Así como las cintas se pierden con el tiempo, se dañan también con el uso. Lo peor es que la copia es siempre es de peor calidad que el original. La calidad de la grabación y los datos lingüísticos quedan comprometidos, y siempre se degradan.

3. Los formatos digitales libres, la solución definitiva.

Los problemas mencionados arriba pueden resolverse con el uso de formatos digitales para grabar, copiar, guardar y distribuir datos, con alta calidad y mayor comodidad de manipulación de los mismos por parte del investigador. Las ventajas principales son:

- Podemos adelantar y retroceder tantas veces como queramos el sonido para escucharlo cuantas veces queramos, sin el riesgo de dañar el medio.

  Al contrario de las cintas magnéticas que pueden romperse durante ese procedimiento, los datos digitales pueden adelantarse y retrocederse sin problemas. Los datos digitales no pierden calidad por este proceso.

- Los datos digitales se degradan muy poco con el tiempo.

  Los datos digitales prácticamente no se dañan con el tiempo. La vida de una cinta DAT, o de un CD-ROM bien cuidados son mas largas que la de una cinta. Además, como las copias recuperan la calidad del original, sacando copias nuevas cada 2 o 3 años y reemplazándolas por los originales aseguraremos preservar los datos tanto tiempo como queramos.

- Existen medios digitales que se resisten a la humedad y a los campos magnéticos fuertes.

La tecnología digital ha estado desarrollando mucho en este últimos seis años y, hoy, nosotros podemos encontrar en el mercado formatos bastantes resistentes, como es el caso de los CD-ROM industriales. Con esos formatos, los datos en ellos almacenados no pierden. Los CDs grabables son mucho más delicados, y no resisten la humedad -aunque la resistan mejor que las cintas de audio-, pero si los campos magnéticos fuertes.

- En espacios pequeños podemos tener grandes cantidades de grabaciones de hablantes.

  En el mundo moderno, la falta de espacio es un problema que nos afecta directamente, sobre todo cuando estamos hablado varias horas de horas de grabaciones para cada hablante, con cientos de hablantes. En un solo CD-ROM, en formato mono -suficiente para un hablante, ya que nos interesa la calidad del sonido, no el estereo- podemos ahorrar el espacio físico de aproximadamente 12 o más cintas cassettes de 60 minutos, dependiendo de tipo de grabación seleccionada.

- Y el más importante: la copia tiene la misma calidad que el original.

  Al contrario de las cintas magnéticas, los datos digitales no pierden su calidad cuando se copian. La calidad se queda así como en el original y, haciéndo copias de seguridad de los datos guardados, estamos seguros que los datos nunca se perderán. Este procedimiento es más simple y mucho más barato económicamente que las limpiezas tradicionales de las cintas cassettes para quitar humedad.

El hecho de que el formato digital sea libre es fundamental si pensamos dentro de una década, o un siglo, cuando no queden hablantes vivos del dialecto estudiado, o sea necesario hacer un estudio diacrónico -estudio de la evolución temporal de un dialecto-. El formato debe ser abierto, para que en el futuro los datos sean legibles por los investigadores, y libres, para que no sea delito construir un reproductor de dichos formatos.

4. Formato digital y medio digital escogido.

Como medio digital hemos escogido el CD-ROM grabable, por su alta capacidad, bajo precio y porque las copias son iguales al original. El problema de ser el CD-ROM grabable sensible a perforaciones, suciedad y humedad se ha resuelto sacando varias

copias de los datos, y guardándolos en lugares distintos. Solo se echa mano de las copias guardadas para sacar copias de uso, con el matiz de que cuando se saca una copia de uso se comprueba la copia, y se guarda la copia en lugar del original y se pasa a usar el original, para asegurar la rotatividad de CD-ROMs. Esta dinámica ha asegurado dos años de uso continuo de gran datos lingüísticos por un grupo de investigación con poca o nula experiencia informática, sin perdida de datos -algo que con el mecanismo antiguo de cintas no era posible-.

El formato digital ha sido un problema más delicado. Cuando comenzamos el proyecto hace dos años tuvimos que escoger MP3, a pesar de ser un formato patentado y no libre, por varias razones: era abierto y la situación de las patentes no había llegado a los niveles actuales -que violan el sentido del ridículo-. Cuando comenzamos a trabajar en nuestro programa que graba directamente de una forma amigable para el lingüista, la capacidad de grabación de una corriente de datos "on the fly" de forma fiable de Ogg Orbis era limitada.

Por ello, hemos grabado muchos datos en formato MP3, primero convirtiendo los datos de los últimos años de investigación a Wav y posteriormente a MP3 usando bladenc. Cuando nuestro programa fué desarrollado, los datos fueron codificados directamente con nuestro programa a formato MP3. Ahora estamos trabajando en portar nuestro programa a Ogg Orbis, para poder liberarlo sin problemas legales. La próxima versión de nuestro programa, del que hablaremos más adelante, soportará Ogg Orbis como formato nativo.

5. El sistema operativo Linux como alternativa para el uso de nuestro software

Es muy común oír las expresiones del tipo: " Linux es muy difícil de usar", "Eso es sólo para el gurú", entre otras. Entornos como KDE han permitido que el sistema sea usado por lingüistas sin problemas de adaptación al nuevo entorno, y con fiabilidad, sin cuelgues, ni pérdida de datos, ni problemas de virus. Todas las aplicaciones usadas, salvo el VARBRUL, tienen un equivalente para Linux, por lo que el tránsito ha sido fácil. Por ello, la opción escogida ha sido Linux+KDE, con un estilo tipo Windows 95. El uso de KDE ha permitido una adaptación automática de los lingüistas al nuevo entorno, siendo poco perceptible para la mayor parte de los usuarios el cambio de sistema gracias al estilo de Windows 95.

Otro problema distinto es el de la instalación de Linux. El hecho de que no sea posible en Brasil comprar máquinas con Linux preinstalado ha hecho que tengamos que

depender de un informático para enseñarnos a instalar Linux y configurarlo adecuadamente. Además, ha habido que escuchar muchas tonterias de los vendedores de hardware cuando algo fallaba dentro de la garantía. Como ejemplo, una vez que el procesador de un ordenador se quemó porque el ventilador no había sido colocado correctanente, la escusa de la tienda para no responder a la garantía fue que el procesador se quemó porque tenía dos sistemas operativos, y "todos saben que con dos sistemas operativos las máquinas se calientan el doble".

La distribución empleada para nuestra investigación ha sido la Mandrake, por su comodidad de instalación y por tener todas las herramientas que necesitábamos en los CDs que pueden ser descargados de Internet gratuitamente. Aunque Debian fue una primera opción, el hecho de no tener un mecanismo de instalación comprensible por un lingüista, y el hecho de no traer KDE por defecto hizo que fracasara el primer intento con Debian por un exceso de dependencia con el informático, y finalmente escogiesemos Mandrake como opción. Cualquier otra distribución razonablemente completa debería ser válida, incluyendo Debian cuando tenga un mecanismo de instalación comprensible por no informáticos.

6. EL proceso de grabación

En un primer paso, teníamos gran cantidad de cintas de investigaciones antiguas que corrían riesgo de perderse. Por ello, digitalizamos todas las cintas con el programa Broadcast 2000. Después convertimos los datos de formato WAV a formato MP3 con bladenc. Estos datos siguen siendo usados en formato MP3 para investigaciones en la actualidad, sin ninguna pérdida asociada al uso continuado por varios investigadores al que han sido sometidos los datos.

Sin embargo, en el proceso de conversión de cinta a MP3 se perdía en calidad, y ello nos llevó a desarrollar un programa, el liverecord. Este programa graba y codifica en formato MP3 en vivo, grabando ya en formato MP3 por lo que podemos grabar horas de audio sin llenar el disco duro, que en los portatiles suelen ser pequeños. Nuestro programa tiene los mismos botones que un grabador tradicional, más dos campos: frecuencia de grabación y tiempo de corte. Cada tiempo de corte el programa cierra el archivo que se está grabando y genera un archivo nuevo, lo que facilitará el uso posterior para organizar los datos. El proceso grabación es simple. El investigador en lugar del grabador y el micrófono puede llevar un portátil y un minimicrófono de solapa, y activar el programa que hemos desarrollado. El resto lo hace el programa solo.

El programa ha sido desarrollado sobre KDE usando Kdevelop. Ahora no está disponible por problemas legales relacionados con el formato MP3 -podemos ser

procesados legalmente si lo liberamos-; estamos trabajando en la conversión del programa a Ogg Orbis -conversión que supone no solo cambiar el formato de grabación, sino también incluir un interfaz amigable para la audición de datos, vease el próximo punto-; en el momento que la conversión sea realizada el programa será disponibilizado en la red. El coautor de este trabajo, que es el informático mencionado en los puntos anteriores, ya no es más becario de investigación y trabaja en la industria privada, por lo que las fechas de terminación están abiertas y dependen de su disponibilidad de tiempo libre.

7. La audición de los datos

Para el trabajo de audición de datos, el programa que adoptamos fue el Broadcast 2000. Es de fácil manejo, y en él hay herramientas que lo hacen indispensable para el tratamiento acústico de los datos, como: demarcación de frecuencia y de niveles, demarcación espacial en la grabación que debe repetirse tantas veces como sean necesarias, y filtros acústicos que permiten limpiar los ruidos, entre otras utilidades.

En total, en los últimos años hemos procesado más de 1500 horas de habla, con una comodidad impresionante.

Sin embargo, el hecho de mover nuestro programa a Ogg Orbis nos va a suponer un problema, ya que el Broadcast 2000 no soporta Ogg Orbis. Los programas que existen para Ogg Orbis están aún muy lejos de lo que necesitamos para nuestra investigación, por lo que en la conversión a Ogg Orbis estamos también desarrollando el interfaz gráfico de audición.

8. Procesamiento de datos

El procesamiento de datos ha sido realizado con el programa VARBRUL, programa especializado en el cálculo de interdependencias de datos fonéticos. Desgraciadamente no hay equivalente libre, por lo que tuvimos que usarlo desde xdos con freedos. No conocemos planes de desarrollo de ningún projecto libre para sustituirlo.

9. Procesamiento de textos

Una vez calculados los resultados, hay que publicarlos en revistas científicas. La mejor

solución para su publicación sería LaTeX, como veremos en los próximos puntos. De hecho, uno de los autores de este texto, lingüista de formación, esta usando LaTeX para redactar su doctorado.

El problema es que ninguna revista de lingüística acepta LaTeX, por lo que hay que adaptar el artículo a formato Word. El único de los procesadores de textos para Linux que exporta a Word y no se cuelga, ni destroza el formato, ni destroza el fichero es StarOffice, que no soporta las fuentes fonéticas, por lo que no hay ninguna alternativa razonable libre a Word que permita exportar a formato Word y soporte el alfabeto fonético.

Este no es solo un problema de la lingüística: también lo encontramos en Linux: las revistas y los congresos para Linux tienen los mismos problemas. La mayor parte de las revistas solo aceptan Word; este mismo congreso solo acepta DocBook, con lo que hemos tenido que aprender otro sistema, mucho menos potente, para poder presentar el trabajo, y que tampoco nos permite soportar el alfabeto fonético, por lo que DocBook tampoco valdría para trabajar para lingüística.

Las razones por la que LaTeX sería perfecto es:

- Calidad profesional y economía del espacio.

  El resultado final de LaTeX para publicar textos, principalmente artículos, relatorios, disertaciones, tesis, etc., es excelente. El producto final tiene calidad profesional, lo que no se consigue con Word.

- La economía de espacio fue otra razón importante. Uno de los autores de este artículo necesitó nueve disquetes para una disertación de mestrado de 135 páginas, y con problemas en las impresiones por las imágenes -con fuerte personalidad-. El otro autor de este artículo realizó un proyecto fin de carrera de casi 500 páginas, con gran cantidad de gráficos y ecuaciones de mecánica cuántica. Cabía en un disquete, y se imprimía en cualquier impresora sin problemas.

10. Calidad en las publicaciones

Uno de los autores de este artículo ha visto como sus artículos eran destrozados al ser publicados en revistas del área. Complejas transcripciones fonéticas en IPA eran convertidas en ristras de olitas, muñequitos y símbolos de Yin-Yan. A este mismo congreso con DocBook habría sido imposible presentar un trabajo de fonética, ya que DocBook no tiene capacidad de representar el IPA.

Por otro lado, con el potentísimo paquete Tipaman de LaTeX podemos realizar transcripciones fonéticas de gran complejidad con sencillez, y con un resultado que no desaparece, los gráficos no saltan, y no depende de la impresora.

Además, otras características de LaTeX interesantes son:

- Los gráficos no desaparecen.
- Los gráficos no se deshacen por la página.
- Si se modifica el texto, el índice se ajusta solo.
- Si se modifica el texto, la bibliografía se ajusta sola.
- Las notas de pie de página están siempre donde deben.
- Si se imprime dos veces el mismo trabajo tiene el mismo número de páginas.
- Se pueden imprimir las 100 primeras páginas en una impresora, y otras 100 páginas en otra sin tener que tocar el texto
- El proceso de impresión es fácil y limpio.
- Los diacríticos salen siempre encima de la letra donde se ponen -el alfabeto fonético, con casi una docena de diacríticos, esto es especialmente importante-.
- Si se manda el texto a otra persona por correo electrónico, él lo imprimirá exactamente como fue generado. Sobre todo, no cambia el IPA por simbolitos raros.

11. Conclusión

Actualmente un lingüista puede utilizar software libre para prácticamente todo el proceso de su investigación, salvo en el estudio de correlación de datos lingüísticos, que se hace con VARBRUL, y en la creación de los artículos para revistas, que se debe usar un WinWord antiguo con Wine. El avance de los editores de textos libres hace suponer que en el futuro el WinWord será prescindible, aunque no soluciona el problema principal: la única forma que hemos encontrado de poder realizar transcripciones fonéticas de calidad con IPA ha sido mediante LaTeX; lo que sirve apenas para tesis doctorales, ya que no hay revistas lingüísticas y prácticamente no hay revistas de informática que acepten LaTeX. Por ejemplo, sería imposible incluir una transcripción fonética como demostración en este congreso.

El uso de Linux es fácil para los lingüistas; salvo en la instalación de Linux, y en la reclamación ante fallos del hardware en garantía, en el que le echarán los montadores de ordenadores la culpa a Linux para no responder de la garantía. Vendedores de

ordenadores que vendan máquinas con Linux preinstalado y que no se escuden en Linux para no responder ante la garantía supone un paso fundamental en este aspecto.

Hemos desarrollado un programa para grabación en vivo, liverecord, para permitir solucionar uno de los problemas de la investigación. Nuestro problema fundamental es que este programa fue desarrollado codificando en MP3, y liberarlo supondría una quiebra de patente. Estamos convirtiendolo a Ogg Orbis -despacio, desgraciadamente, ya que el que lo está haciendo lo hace en su escaso tiempo libre-. Esperamos mejorar este programa y poder desarrollar más código de lingüística computacional en el futuro.

Maria Francisca Ribeiro de Araujo tiene una beca de doctorado de la FAPESP.

Bibliografía

Carlota Ferreira y Suzana Cardoso, *A dialetologia no Brasil: Metodologia do trabalho lingüístico e atlas dialetológico, regionalismos léxicos*, 1a edición, Contexto, 1994.

William Labov, *Principles of linguistic change: Internal factors.*, 1a edición, Blackwell , 1994.