TshwaneLex, una aplicación lexicográfica singular

Ignacio Navascués Benlloch*

Resumen: TshwaneLex es una *suite* excelente compuesta por una base lexicográfica de datos, un programa de gestión terminológica y un 'lector', y reúne casi todos los medios necesarios para elaborar un buen diccionario monolingüe, bilingüe o multilingüe. La interfaz sencilla, sobria y estable de la aplicación facilita un uso cotidiano intuitivo, que destaca por la facilidad para crear y visualizar el artículo, la presencia de una formidable ventana de trabajo con la posibilidad de acometer todo tipo de consultas y búsquedas, la gestión y el rastreo de las referencias cruzadas y la importación parcial o completa de la obra con distintos formatos. La administración de ciertas funciones de enorme relevancia, como la configuración de la plantilla, la fusión de bases con fechas distintas, la importación de ficheros o el trato de las formas complejas, exige del usuario conocimientos más que elementales.

Palabras clave: lexicografía, terminología, aplicación informática, diccionario, glosario, microestructura, referencias cruzadas, trabajo en red, formas complejas, suano, TshwaneLex.

TshwaneLex, a unique lexicographic application

Summary: TshwaneLex is an excellent suite comprising a lexicographical database, a terminology management program and a 'reader'; it brings together almost all the necessary resources to prepare a good monolingual, bilingual or multilingual dictionary. The day-to-day use of the application is comfortable and intuitive from a simple, sober and stable interface that stands out for its ease in creating and viewing the article, the presence of a magnificent work window with the possibility of carrying out all kinds of searches and filters, the management and tracking of cross-references and the partial or complete importation of work in different formats. The administration of certain functions of enormous relevance, such as the configuration of the template, the merging of bases with different dates, file import or the handling of compound lemmas, requires more than elementary expertise on the part of users.

Key words: lexicography, terminology, computer application, dictionary, glossary, micro-structure, cross-referencing, networking, compound lemmas, *suano*, TshwaneLex.

Panace@ 2009; 10 (29): 93-97

1. Origen de la *suite:* el programa lexicográfico

En Tshwane, más conocida como Pretoria, nació en el año 2003. TshwaneLex, una aplicación para la compilación inicial de diccionarios y glosarios a la que he bautizado como *suano.* Sus creadores, David Joffe y Gilles-Maurice de Schryver, han logrado entre tanto que esta aplicación, única en su género, sea utilizada en más de 150 lenguas diferentes y por más de 300 clientes del mundo entero, entre los que sobresalen la Real Academia Española, la Real Academia Nacional de Medicina, el Instituto de Lexicología Holandesa, el Instituto de Lengua y Literatura de Malaisia, el Centro de Investigación de Lenguas Africanas, el Instituto de Investigación Científica y Tecnológica de Ruanda, la Academia de Ciencias de Eslovaquia, el Instituto de la Lengua de la República Checa, el Departamento de Justicia de Canadá o las editoriales Grupo Clarín, Oxford University Press, Le Robert, Van Dale Lexicografie, MacMillan y Pharos Dictionaries.

2. Programas lexicográficos

En el mundo de la lexicografía se oyen burlas frecuentes acerca del programa más utilizado en la compilación de los diccionarios. No es otro sino el famoso Word de la multinacional Microsoft, que no nació precisamente para crear diccionarios. Lo más parecido a una base de datos de Word son sus tablas. Los lexicógrafos, gente paciente, esforzada, meticulosa y puntillosa donde las haya, llevan lustros soportando la carestía de recursos informáticos. En el mejor de los casos, imploran a algún experto informático que les diseñe un programa a medida, y en el peor, las editoriales les reclaman la elaboración de pequeños programas o macros para su incrustación dentro de Word. Los programas lexicográficos constituyen una especie rara.

3. Requisitos de un programa lexicográfico

3.1. Asistir en la creación de la planta del diccionario, fase inicial ardua, que requiere largo tiempo. La macroestructura se extrae, de ordinario, de corpus ya existentes, mientras que la microestructura sienta las diferencias entre unos diccionarios y otros. Dotar de vida al artículo precisa bastante habilidad. Ayudar a su gestación desde el propio *software* y trasladar luego la estructura subterránea a la superficie allanan considerablemente la labor lexicográfica.

3.2. Gestionar las tareas de los colaboradores del diccionario.

3.3 Trabajar en equipo mediante una red física o virtual o, al menos, reunir los productos de cada uno para anexionarlos al embrión lexicográfico.

* Médico traductor, Madrid (España). Dirección para correspondencia: ignacio.navascues@telefonica.net.

3.4. Disponer de una interfaz visual cómoda, intuitiva y fácil de utilizar.

3.5. Consultar la base de datos para detectar errores y omisiones; administrar los lemas en sus diferentes fases de elaboración; seleccionar entradas que requieran un trato especial; dar unidad a lemas afines, etcétera.

3.6. Administrar las referencias cruzadas entre lemas, posiblemente uno de los caballos de batalla más duros de pelar. Hasta no hace tanto tiempo, y sin duda todavía en muchos lugares, las referencias se colocaban o se siguen colocando a mano o con macros de un procesador de texto u otros sistemas complicados. Sin embargo, su naturaleza y orden cambian con frecuencia a medida que se van modificando, ampliando o eliminando vocablos del diccionario. Si no se dispone de un editor adecuado, que impida la asignación de referencias a lemas inexistentes, la invitación al caos parece servida.

3.7. Administrar las formas complejas de los lemas simples. En principio, las formas complejas tendrían que acomodarse dentro de las simples sin que la base de datos ponga trabas para organizar la información.

3.8. Decidir el formato final de la obra. No es lo mismo imprimir un diccionario que preparar una versión electrónica o en línea para su consulta desde un ordenador o a través de la red, respectivamente.

4. Características de la suite TshwaneLex (<http://tshwanedje.com/>)

4.1. Versiones y precios

Desde el año 2007, y a partir de la tercera versión de TshwaneLex, esta *suite,* aparte de la base lexicográfica inicial de datos, incluye una herramienta terminológica, TshwaneTerm, y un lector gratuito, TshwaneReader, para las personas sin licencia. La compañía ofrece dos tipos de licencias, una comercial por 1900 euros y otra académica por 150 euros.

La *suite* está basada en el lenguaje XML y soporta Unicode. Las versiones actuales son compatibles con Windows XP y Vista. El *suano* dispone de una interfaz de programación (API) en lenguaje Lua para el gobierno íntegro de la base de datos.

4.2. Tamaño de la base de datos

El peso de la base de datos depende de la microestructura y de la macroestructura. Mientras Word se arrastra cual tortuga con documentos de tan solo 10 MB, TshwaneLex gestiona bases de 80 MB con gran agilidad. La velocidad depende también del microprocesador y de la memoria de la computadora.

4.3. Tipos (proyectos) de diccionarios

La base de datos permite iniciar proyectos para diccionarios bilingües, monolingües y multilingües. El *suano* ayuda sensiblemente a la confección de diccionarios bilingües, ya que dispone de herramientas que revierten los lemas de un idioma a otro a través del elemento llamado «equivalencia» (TE, *translation equivalent*).

4.4. El artículo suano y su (micro)estructura

4.4.1. La plantilla tldtd

El programa trae, por suerte, una plantilla predeterminada, basada en la DTD (definición del tipo de documento) del lenguaje XML, que dispone de los requisitos básicos para iniciar cualquier proyecto. La primera misión, nada fácil, consistirá en adaptar esta microestructura. Una vez configurada, el administrador deberá guardarla (en formato *.tldtd) y «retocarla» lo menos posible. La plantilla se puede copiar a otros proyectos, y su formato difiere del de la base lexicográfica (*.tldict).

4.4.2. Nivel del usuario y usuarios de nivel

En la propaganda de la compañía se lee, a modo de gancho, que esta *suite* puede ser utilizada por personas sin conocimientos informáticos profundos, y resulta, en gran medida, cierto. Sin embargo, para adentrarse con éxito en la microestructura y para realizar alguna que otra tarea de mantenimiento, como la fusión de diccionarios o la importación de ciertos documentos, se requieren conocimientos medios o avanzados.

4.4.3. El árbol, los elementos y los atributos

El artículo *suano* se parece a un árbol con un *tronco,* el lema; *ramas,* los demás elementos, casi todos configurables por el usuario, y *hojas,* los atributos, donde se almacenan los datos, los colores, los estilos, las definiciones y todos los demás valores.

4.4.4. Los elementos

Es posible añadir cualquier elemento y cambiar su nombre o su orden de aparición. Conviene separar el *elemento* del *atributo,* siempre subordinado al primero. Cada elemento puede tener dos clases de hijos: los atributos y otros elementos. El número de hijos lo define el usuario.

4.4.5. Los atributos

El atributo *suano*, la materia prima del usuario, consiste en un valor fijo, único, múltiple, variable, alfabético o numérico, una imagen, un sonido o incluso un guión en Lua. Las «listas de atributos», opción utilísima, permiten colocar latiguillos como «evítese», «suscita rechazo», «v.», «s. m.» mediante un simple clic. Los atributos también se pueden cambiar de nombre y de orden.

4.5. Orden de visualización (salida) y formato de los atributos

El orden de visualización de los elementos y sus atributos se modifica desde la pestaña «Output (display) order» de la plantilla. Con la pestaña «Styles/formatting» se maqueta el artículo (atributos de letras, sangría, interlineado, etc.).

5. La interfaz visual *suana*

5.1. Aspectos generales

Cuando se abre por primera vez la aplicación y se inicia un nuevo proyecto, surge una ventana que solicita la información básica. Al presionar el botón «Ok», se despliega la interfaz de la aplicación; comentaré solo la ventana de los proyectos monolingües.

5.2. Ventanas

El lexicógrafo trabajará con una interfaz relativamente fija, compuesta por cuatro ventanas, que describiré de izquierda a derecha (figura 1).

5.2.1. La ventana del lemario

La estrecha ventana vertical del lemario se divide en tres secciones desiguales bajo unos botones. La primera es una casilla para la consulta rápida de los lemas. La segunda, más extensa, muestra el lemario alfabetizado. En la sección inferior, breve, figuran los vocablos consultados en cada sesión, que se pueden rescatar en cualquier momento.

5.2.2.La ventana del árbol o de los elementos

La ventana del árbol, en la mitad central superior, permite configurar los elementos del artículo. Los elementos del lema se pueden subir y bajar o copiar, cortar y pegar con suma facilidad dentro del artículo. Parte de esta labor se automatiza con ficheros en formato CSV.

5.2.3. La ventana de trabajo o F1 o de los atributos y útiles

La ventana real de trabajo, en la mitad inferior central, se compone de seis subventanas, de F1 a F6, que se abren con la tecla de función correspondiente. Las únicas ventanas para la introducción de datos son F1, F2 y F12.

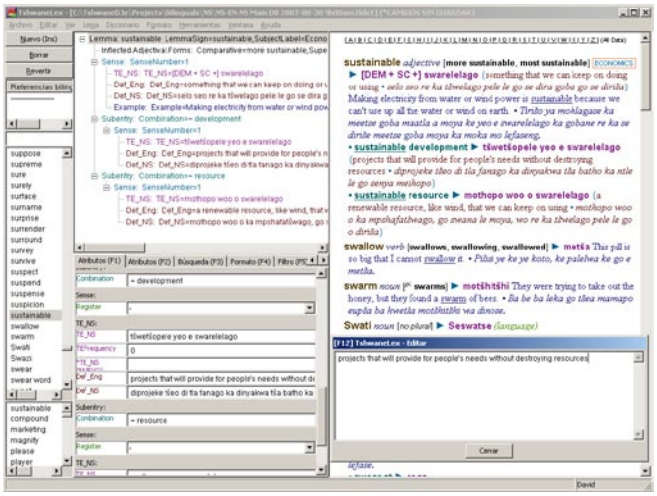

Figura 1. Interfaz de trabajo de TshwaneLex, en la que el marco flotante F12 aparece incrustado en la parte inferior de la ventana de previsualización. (Imagen tomada de un diccionario inglés-sotho del norte y reproducida con permiso de David Joffe)

F1 aloja los datos en casillas de atributos y, además, muestra el valor de las listas de una sola opción. Con F2 se seleccionan los ítems de las listas múltiples, lo que ahorra el tecleo de textos iterativos y asegura su uniformidad. F3 busca datos y, si no dispone de espacio suficiente, recurre a «Filter», desplegando todas las entradas filtradas en los demás marcos. La búsqueda se puede acotar por campos o mediante expresiones regulares. La opción de reemplazo se activa, curiosamente, desde una pestaña del menú principal. F4 modifica el formato de la ventana de previsualización. F5 filtra (consulta); dada su importancia, expondré las consultas más adelante. F6 agrega corpus desde cualquier almacén del ordenador e incorpora fragmentos de texto al elemento «Example».

La ventana de trabajo no resulta la más adecuada para la confección de enciclopedias o diccionarios con una extensión moderada; para ello sirve, en cambio, el marco F12.

5.2.4. La ventana de previsualización o del artículo

Esta ventana ocupa la mitad derecha de la pantalla y muestra el artículo completo tal y como haya sido definido (WYSWYG) en la plantilla. Las casillas de F4 controlan la visualización de hipervínculos, lemas con referencias cruzadas al artículo o fechas de edición.

5.2.5. La ventana mágica de edición F12

El *suano* perdería mucho atractivo lexicográfico si no dispusiera de la ventana autónoma F12, de extraordinaria utilidad para la redacción y edición de textos largos, así como para el pegado de otros similares.

6. Primeros pasos con TshwaneLex

6.1. Creación de la plantilla

Véase el apartado 4.4.1.

6.2. Creación (y borrado) de lemas

Los lemas se crean con la tecla Ins o el botón «New (Ins)» y se pueden borrar con una combinación de teclas o con el botón «Delete».

6.3. Creación del artículo

La forma del artículo variará de un lema a otro; el lexicógrafo pasará muchas horas en la ventana del árbol saltando de rama en rama y configurando elementos.

6.4. Cumplimentación de los datos del artículo

En esta etapa, la más entretenida y gratificante, se emplean las ventanas F1, F2 y F12. El *suano* escribe bien cualquier carácter y facilita el reemplazo automático de cadenas de texto, pero jamás muestra en F1, F2 o F12 atributos como el subrayado o los colores. Estos solo se ven en la ventana del artículo.

7. Trabajo en equipo

La aplicación permite trabajar en una red física por medio de servidores ODBC. La labor de cada usuario se incorpora al instante.

7.1. Colaboración desde lugares remotos

Como la mayoría de los colaboradores trabaja en lugares distintos y distantes, el equipo lexicográfico debe plantearse la inclusión en la plantilla de elementos y atributos para la administración y gestión. Los diseñadores del *suano,* conscientes de esta demanda, ofrecen reunir en un documento central los productos elaborados por los diferentes colaboradores.

7.1.1. Comparación y fusión de los embriones lexicográficos

Con la opción «Compare/Merge dictionary» del menú principal se compara el diccionario abierto en ese momento

con cualquier otro. El resultado se visualiza en una ventana nueva.

Los lemas del segundo diccionario se pueden incorporar al original abierto, no hay otra alternativa, de tres maneras diferentes: adición («Add»), mezcla («Merge») y reemplazo («Replace»). Cabe también la posibilidad de borrar lemas del diccionario central («Delete left»).

Los lemas se pueden fusionar uno a uno y también «por lotes» («Batch merge»). Antes de la fusión, conviene cerciorarse de que los lemas visualizados en la pantalla de comparación sean los adecuados y de que no falte ni sobre ninguno. Esta operación, la más delicada de todas, no puede iniciarse sin guardar antes una copia de seguridad del proyecto.

8. Las referencias cruzadas

Una de las cualidades más sobresalientes de esta aplicación es el manejo de las referencias cruzadas. Aparte del editor, donde se crean todas las variantes imaginables (figura 2), el punto fuerte radica en la gestión: *a*) no se puede dirigir una referencia a un lema inexistente, lo que impide su extravío; *b*) el registro central lleva cuenta de cada referencia a la acepción o al lema asignados; *c*) cuando se borra un lema con referencias, la aplicación advierte de la eliminación de estas últimas; *d*) las referencias actúan como hipertexto; *e*) los lemas referenciados se visualizan, F4 mediante, en la ventana del artículo.


haria ♦see hariat

Este artículo envía referencias a >>

hariat haria, aria, ariasse [harja,hărjă, arja, arjas] n.m. 1 good-for-nothing person ▸ *Le monde qui reste auprès de l'école, c'est des harias.* The people who live near the school, they're trash. (SM)
2 *(often pl.)* trash, anything old or beat up ▸ *J'ai nettoyé ma grande closet et j'ai porté un tas d'hariats dans le garage.* I cleaned out my big closet and I took a pile of trash into the garage. (Lv88) ▸ *C'est un tas d'hariat.* That is a lot of trash. (JE) ▸ *Ce capot est un haria.* That coat is an old rag. (IB)
3 bother, row ▸ *Quel aria!* What a bother! (SM)
<*Loc:* EV, IB, JD, JE, LA, LF, SB, SJ, SM, TB, Da84, Di32, Lv88>

>> Este artículo recibe referencias de

aria, ariasse ♦see haria

Figura 2. Ejemplo de referencia cruzada tomado de un diccionario de francés de Luisiana. (Reproducido con permiso de David Joffe)

La latencia de apertura de la ventana de asignación de referencias se antoja excesiva o, en las enciclopedias o diccionarios con más de cien mil entradas, prohibitiva.

9. Consulta de la base de datos o ventana de filtro F5

El *suano* posee una subventana prodigiosa de consulta, F5, con dos columnas verticales donde se disponen todos los elementos y atributos con una sencillez que impone una sensación inicial de extrañeza. En la primera columna se marcan los valores que se quieren filtrar y en la segunda, los que se desea excluir, combinados o no con los operadores booleanos *O* e *Y*. El marco F5 será muy apreciado tanto por los administradores como por los lexicógrafos.

10. Importación y exportación de datos

La aplicación lexicográfica brinda a sus clientes la opción de importar datos en lenguaje XML y en ficheros con formato CSV y de exportarlos con diversos formatos. Además, se ofertan módulos para la confección de diccionarios electrónicos y en línea.

La exportación es la operación más sencilla. La importación constituye, en cambio, una de las utilidades artesanas más robustas y delicadas y no debe acometerse sin guardar antes una copia de respaldo. Conviene recordar que los ficheros de texto exportados o importados se almacenan con la codificación UTF-8 y que esta no debe, en principio, cambiarse.

El mayor inconveniente de la importación de ficheros CSV reside en la imposibilidad de asignar valores a acepciones de lemas polisémicos. El engorro para organizar la importación de los datos de distintas acepciones en formatos CSV es tal que ni siquiera vale la pena el intento. Con los homónimos ocurre lo mismo que con los lemas polisémicos.

11. Las formas complejas

El tratamiento de las formas complejas de los lemas está sujeto, como en muchos diccionarios, a la virgulilla (~). Las formas complejas se pueden visualizar sincopadas (con virgulilla) o expandidas; de ello se ocupa F4. Si el usuario desea compilar un diccionario de extensión mediana o grande, encontrará bastantes trabas para diseñar, visualizar, editar y revisar las formas complejas, pues la ventana del árbol no puede mostrar más allá de seis acepciones, por lo que el lexicógrafo acabaría hartándose de bajar y subir la barra de desplazamiento. Para la elaboración correcta de los lemas compuestos con el *suano,* hay que darles el mismo trato que a los simples.

12. Impresión general del trabajo con TshwaneLex

Esta aplicación brinda un útil excelente al usuario común para la compilación lexicográfica con una interfaz intuitiva, sobria, cómoda, sencilla y sumamente estable, dotada de casi todos los instrumentos necesarios. No obstante, se necesita un administrador de la base lexicográfica para ciertas operaciones esenciales, como la creación de la plantilla o la fusión de documentos.

13. Guía para el usuario y asistencia técnica

La guía del *suano* ayudará al usuario a dar sus primeros pasos. La asistencia técnica por correo electrónico es competente, seria, rápida y cordial.

14. Consejos prácticos de uso

- Adquirir un buen ordenador.
- Fechar a diario el documento central y llevar un registro seguro.
- Comprobar con regularidad los errores.
- Diseñar con esmero la microestructura y retocarla lo menos posible.
- Crear elementos antes que atributos.
- Incorporar los latiguillos lexicográficos a las listas de atributos.

- Incluir elementos y atributos de administración y gestión, si no se trabaja en una red física.
- Abrir varios documentos suanos para evacuar consultas simultáneas.
- Usar la ventana F12 para textos extensos.
- Utilizar la función «Unapply» para buscar e intercalar lemas durante las consultas con F5.
- Armarse de paciencia para las operaciones de fusión y de importación.

15. Propuestas finales

La aplicación lexicográfica TshwaneLex Suite es muy valiosa y se perfeccionará, sin duda, en el futuro. Para terminar, me gustaría formular algunas propuestas a los autores:

- Ofrecer una licencia de bajo coste para terminólogos, traductores y lingüistas que únicamente quieran elaborar proyectos caseros.
- Acortar la latencia de apertura de la ventana para la asignación de referencias cruzadas.
- Plantear la opción de independizar la acepción del lema para darle un trato parejo y simplificar así la introducción de valores iterativos desde los ficheros importados.
- Segregar los lemas monosémicos de los polisémicos, si no se resuelve el punto anterior.
- Replantear el tratamiento de las formas complejas.
- Resaltar con hipertexto las entradas del lemario en los diferentes campos.
- Incluir un corrector ortográfico, quizá a modo de extensión.

Agradecimientos:
Deseo agradecer las sugerencias formuladas por mis colegas María Teresa Sánchez Safont y Fernando Navarro tras la lectura del borrador.

Darwin y el español

Fernando A. Navarro

Este año 2009 viene marcado por una doble conmemoración darwiniana: el pasado 12 de febrero se cumplieron exactamente doscientos años del nacimiento, en la localidad inglesa de Shrewsbury, del naturalista Charles Darwin, y el próximo 22 de noviembre se cumplirán exactamente ciento cincuenta años de la publicación de su obra científica más destacada: *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life.*

Aprovechando la ocasión, me gustaría señalar un pequeño pasaje de otra obra suya mucho menos conocida. En los apuntes autobiográficos que Charles Darwin escribió en 1876 —publicados de forma póstuma en 1887 en *The life and letters of Charles Darwin, including an autobiographical chapter*—, el naturalista inglés comenta en los siguientes términos la fabulosa acogida que obtuvo su *Origin of species,* éxito editorial desde el primer día:

> It is no doubt the chief work of my life. It was from the first highly successful. The first small edition of 1250 copies was sold on the day of publication, and a second edition of 3000 copies soon afterwards. Sixteen thousand copies have now been sold in England; and considering how stiff a book it is, this is a large sale. It has been translated into almost every European tongue, even into such languages as Spanish, Bohemian, Polish, and Russian.

Hace ya tiempo que leí estas notas autobiográficas de Darwin, pero confieso que todavía siento ese «even into such languages as Spanish…» clavado en mi corazoncito como una espina, y cada vez que lo recuerdo me duele profundamente nuestro atraso científico de entonces y de ahora.

De ahora, sí, como puede comprobarse en el mejor cibersitio darwiniano de toda la multimalla mundial, fruto de siete años de duro trabajo por parte de un joven profesor de historia de la ciencia en la Universidad de Cambridge: John van Wyhe. En <http://darwin-online.org.uk/> encontramos recopiladas las obras completas de Charles Darwin: todas sus publicaciones impresas, pero también sus escritos inéditos, 20 000 cartas particulares, una exhaustiva bibliografía darwiniana, el catálogo de manuscritos y centenares de documentos complementarios: biografías, notas necrológicas, reseñas, obras secundarias, etc. En conjunto, más de 70 000 páginas de texto digitalizado (con posibilidad de búsqueda sencilla o avanzada), 175 000 páginas electrónicas facsímiles, tres millares de ilustraciones e incluso una docena larga de audiolibros en capítulos MP3. Y, como era de esperar, el sitio encandila por su exhaustividad en lo que respecta a las publicaciones en inglés, desde luego, pero encontramos también textos en alemán, danés, francés, holandés, italiano, noruego y ruso. Nuestro español, en cambio, ¡ay!, brilla como de costumbre por su ausencia, en clara prueba de cuán poco ha variado nuestra situación al cabo de siglo y medio.