La confección de léxicos específicos para la red. Un ejemplo concreto: el léxico multilingüe de economía del proyecto ElektraVoc-II.

Kris Buyse
en nombre del equipo ElektraVoc

1. Introducción

Para la lectura, redacción o traducción de textos altamente especializados, constituyen un instrumento indispensable los *glosarios especializados* tanto mono como multilingües, en particular los que versan sobre *áreas expuestas a una evolución vertiginosa* como la banca, la nueva economía, etc. Por otro lado, la probabilidad de dar con un término "reciente" en el glosario depende de las *posibilidades de actualización del soporte*, y es sobre todo en este contexto donde empieza a ser útil la confección y disposición de un glosario electrónico 'en línea'. Asimismo, la publicación en formato electrónico, 'en línea' o no, permite añadir a la(s) base(s) de datos otros componentes como un módulo de ejercicios con retroacción, para que el conjunto, además de ser *obra de referencia* altamente accesible y flexible, pueda convertirse en un *entorno de aprendizaje* (semi-autónomo).

No obstante, para evitar que se caiga en las *trampas de los cálculos probabilísticos*, la automatización de (parte de) la labor léxico y/o terminográfica no deja de requerir que un *experto supervise* y controle las fases preparatorias.

En este artículo, voy a presentar, en nombre del equipo ElektraVoc y a partir de los resultados provisionales del módulo económico del *proyecto ElektraVoc-II*[1] *(EV2EC)*, unas pistas y trucos para la elaboración de un glosario económico *on line*.

2. Pistas para la elaboración de un glosario multilingüe "on line"

2.1. Enseñanza y aprendizaje en tiempos electrónicos

Desde finales de los años noventa, caracterizados por la mayor cantidad y disponibilidad de conocimientos e instrumentos asequibles, en contraste con la disminución de las horas de contacto y la imposibilidad de llegar a un conocimiento omnisciente de

1 Colaboraron en el proyecto de vocabulario general *ElektraVoc-I*, subvencionado por la K.U.Leuven (KUL), la Lessius Hogeschool (LHO) y el ministerio de Educación de la Vlaamse Gemeenschap (la comunidad flamenca): Nicole Delbecque (KUL), directora; Kris Buyse (LHO y KUL) y Dirk Speelman (KUL), codirectores; Sara Six e Hilde Hanegreefs (KUL) y Sabela Moreno (LHO), colaboradoras.
El ministerio de Educación también subvenciona el proyecto de vocabulario específico *ElektraVoc-II*, en el que colaboró en el 2002, aparte de la LHO y la KUL, la R.U.Gent (RUG). Además de la directora y los codirectores de ElektraVoc-I, el equipo de ElektraVoc-II constó de los siguientes miembros: Eugeen Roegiest (RUG), codirector; Lieve Behiels (LHO), codirectora para el módulo económico; Katleen Van den Steen (KUL), Patricia Muñoz Tavira (LHO) y Anne Vandenberghe (RUG), colaboradoras. Nicole Delbecque coordina el módulo cinematográfico, Kris Buyse el económico y Dirk Speelman el tecnológico.
Para la ampliación del módulo económico, la LHO y la KUL -cuyos equipos continuaron las investigaciones en el 2003- cuentan a partir de octubre del 2003 con la competencia terminográfica de un nuevo socio, a saber, la Vlekho (coordinador: Danny Masschelein; colaborador: Steven Dewallens).
Para más información sobre el módulo económico: kris.buyse@arts.kuleuven.ac.be.

cualquier rama de la ciencia, vienen imponiéndose poco a poco *nuevos entornos de aprendizaje* de la lengua: en vez de dirigir un transporte unidireccional de conocimientos cerebrales hacia el alumno, el profesor actual se convierte en 'coach' de unos procesos interactivos, encaminados a instruir a los "usuarios" en el *autoaprendizaje asistido* de estrategias y *actitudes* de autoformación 'para toda la vida'.

En estos tiempos, una plataforma electrónica constituida por una base de datos temática de los términos más frecuentes de un área de conocimiento especializada, acompañados por sus fichas terminológicas, introducciones temáticas, ejercicios y una asistencia graduada, permitirá al usuario superar las dificultades que le suelen impedir adentrarse en el área en cuestión. A la base de datos en papel, la electrónica 'en línea' le gana en términos de potencial de accesibilidad, actualización, representatividad, volumen y multifuncionalidad (obra de referencia y al mismo tiempo entorno de aprendizaje). Además, gracias a la *transgresión de la linealidad*, la transparencia *de los criterios y fuentes* de la base de datos está sólo a unos pocos hipervínculos. Asimismo, esta funcionalidad permite integrar mejor en la misma ficha de los lexemas las *relaciones paradigmáticas* (sinónimos, antónimos, parónimos, palabras afines...) y *sintagmáticas* (colocaciones, actantes, ejemplos auténticos...): ambas nos parecen indispensables en la adquisición del vocabulario, que entendemos como el *almacenamiento de un corpus flexible de lexemas y colocaciones básicos, de tal manera que se puedan recuperar fácilmente para su uso correcto y variado.*

EV2EC se enmarca en esta concepción de la confección del saber propio: nos proponemos integrar, en una plataforma electrónica de referencia y de aprendizaje, un corpus representativo de textos económicos, una base de datos terminológica multilingüe con las fichas de sus lexemas y colocaciones más típicos, y un conjunto variado de ejercicios con retroacción, a fin de capacitar al usuario para descodificar artículos económicos en la prensa general y especializada y, al mismo tiempo, invitar a unas primeras actividades productivas en lengua española para fines económicos.

2.2. Selección y confección del corpus virtual

Para el corpus de EV2EC *seleccionamos* unos 5.500 artículos, procedentes tanto de la prensa general como de la prensa económica especializada, ambas 'en línea', y tanto la española como la latinoamericana. Dado el carácter desigual de la cantidad de artículos de índole económica que aparecen en la prensa general, por un lado, y en la propiamente económica, por el otro, fuimos coleccionándolos diariamente, procurando mantener un equilibrio entre las diferentes fuentes.

2.3. Clasificación temática

Para estructurar el léxico económico en sus varias subtemáticas, recurrimos a diferentes *fuentes taxonómicas*. Las clasificaciones de la prensa general y la especializada, completadas por la de los manuales y glosarios de economía y la que manejan las facultades e institutos de economía, nos han llevado finalmente a clasificar nuestros textos en las siguientes subtemáticas : 1) empresa, 2) comercio y *marketing*, 3) recursos humanos, 4) finanzas, 5) economía general y 6) nueva economía:

	TOKENS
1: Empresa (2.21 MB)	378.695
2: Comercio y *marketing* (233 KB)	38.260
3: Recursos Humanos (638 KB)	105.351
4: Finanzas (3.73 MB)	652.902
5: Economía general (5.59 MB)	964.742
6: Nueva economía (1.56 MB)	267.822
Total:	**2.407.772**

cuadro 1: la clasificación temática

Como nuestro objetivo es guiar a los usuarios de la plataforma en su contacto real con la prensa económica, hemos mantenido las desproporciones en el corpus y en las subtemáticas. En el apartado siguiente (2.4) se explica por qué se construye también un subcorpus de economía general (temática 5).

2.4. Extracción, selección y lematización de los términos económicos

En la etapa siguiente, se confeccionó con el programa WordSmith Tools[2] una *lista de palabras* ('wordlist') a partir de los textos sometidos a análisis. A fin de limitar la cantidad de palabras inútiles, trabajamos con una lista de palabras comodines y términos falsos (*'StopList'*), es decir, una lista de palabras que suelen ser frecuentes pero sin ser términos económicos, como términos gramaticales, nombres propios de empresas, etc. Después, se *lematizó* el resultado, agrupando las ocurrencias de un mismo término con sus distintas formas. Por medio del 'KeyWords-tool' de WordSmith, que permite contrastar un subcorpus con el corpus entero o con un corpus de referencia, llegamos a una lista de *palabras clave* ('Keywords'), es decir, palabras típicas para el subcorpus. El programa calcula la *calidad de palabra clave* ('Keyness'), es decir: la medida en que el término caracteriza al texto, teniendo en cuenta las frecuencias relativas de los términos. Los términos económicos con mayor frecuencia –con respecto a unos criterios cuantitativos preestablecidos, a saber: {test=log likelihood, p = 1, max. wanted = 16000, min freq = 1}[3] - pueden, en este sentido, ser considerados como los 'básicos' para el usuario que quiera familiarizarse con el funcionamiento de la economía tal como se percibe y comenta en la prensa.

Durante la misma operación, el Concord-tool permite generar rápidamente todas las *concordancias* del término. Los contextos así obtenidos pueden conservarse en un

2 Programa diseñado por Mike Scott y publicado por Oxford University Press. Unas posibles alternativas son Webcorp, Abundantia Verborum, Trados, ...

3 El test compara un corpus con otro y de esta manera aísla los términos que, según un parámetro de probabilidad, serían "típicos" del primer corpus (i.e. marcadamente más o menos frecuentes que en el otro corpus). Así comparamos los corpus de las subtemáticas económicas (como Finanzas) con el corpus económico general (temática 5) para sacar los términos clave de aquellos corpus. El corpus económico general se ha comparado, gracias a la colaboración de G. Rojo, con un corpus muy amplio de lengua general, a saber, el CREA. Para más información sobre el test, véase Ted Dunning, 1993. Accurate Methods for the Statistics of Surprise and Coincidence. Computational Linguistics 19, 1.
"P" es el parámetro de probabilidad: cuanto más bajo, más relevante será la diferencia. Un valor inferior a 0.05 se considera como significativo. Aquí se ha manejado un valor de 1 para poder comprobar en un orden decreciente todos los "tokens" y filtrarlos manualmente.
Los valores máximo y mínimo de tokens constituyen otro filtro: así se podría limitar la cantidad de términos obtenidos. Sin embargo, aquí se ha optado por manejar unos márgenes muy amplios para poder analizar primero la lista antes de filtrarla manualmente.

documento aparte sin perder las fuentes de procedencia respectivas. Cada contexto se acompaña automáticamente del código que había sido atribuido al documento del que procede. De esta manera, se pueden elegir para cada término unos contextos que apoyan adecuadamente su significado, para uso futuro en las fichas terminológicas del léxico y además para la elaboración del módulo de ejercicios sobre el léxico económico. Asimismo, el *concord-tool* permite aislar las colocaciones más frecuentes. Veremos más adelante[4] cómo se pueden evitar algunas trampas a la hora de aislar las colocaciones.

2.5. Confección de una base de datos

Así seleccionados y clasificados, los términos forman el núcleo central de una base de datos, configurada con la ayuda del programa Multiterm de Trados[5] . Para cada término económico se confecciona *una ficha* en la que, además de la *definición* y la *traducción* del término al neerlandés, francés e inglés, se describe parte de sus relaciones *sintagmáticas y sintácticas*: la categoría gramatical, las colocaciones más frecuentes, los actantes indispensables, unos contextos auténticos, sinónimos, antónimos, parónimos y su frecuencia en el corpus.

Además, gracias a unos enlaces a la bibliografía, pueden consultarse el artículo y el periódico del que proceden, junto con la fecha de publicación en Internet así como las fuentes consultadas para confeccionar las definiciones. Para ello, se elaboraron listas de *fuentes lexicográficas y textuales 'en línea'*, como las siguientes:

1: Español (España) **1*: Español (Latinoamérica)**
2: Nederlands **3: Frans** **4: English**
5: Alemán – Duits **6: Italiano – Italiaans** **7: Pt**

http://labs.google.com/glossary (consulta de definiciones)

1: **Términos económicos** (Corfinsura) [GUESP13]
1: **Vocabulario básico económico** [VUESP1]
1: Glosario de algunos **términos económicos y financieros** [GUESP4]
1*: **Economía**: glosario (México) [GUESP3]
1*: **Glosario económico & financiero** (Colombia)[GUESP12]

1 – 2:
Glosario del euro [GMT1] (2, 3, 4 + resto lenguas comunitarias)

1 – 3:
Banking Terms (1-3-4) [G134T1]
Glosario terminológico sobre **aduanas** (1-3-4-5) [G1345T3]
OECD Macrothesaurus Chapter Headings (4-1-3) [G413T2]
Glosario del **euro** [GMT1] (2, 3, 4 + resto lenguas comunitarias)
Terminología del **FMI** – Glosario multilingüe (1-3-4-5-7) [G13457T3]

cuadro 2: ejemplos de listas de instrumentos lexicográficos 'en línea'

4 Véase el apartado 3.

4 Se podría recurrir a cualquier programa de este tipo, con tal de que el output sea convertible a XML (porque éste permite presentar y utilizar la base de datos de muchas maneras diferentes, sin tener que adaptar siempre el formato).

3. Oportunidades de la plataforma electrónica

Confeccionada de esta manera y publicada tal cual en la web –por medio de la versión iX de MultiTerm-, la base de datos se convierte en una *obra de referencia representativa* y fácilmente *accesible y actualizable*. Sin embargo, se puede ir más allá, aumentando su *representatividad y funcionalidad*, aunque para evitar *escollos* en la ruta hay que contar con la experiencia del *experto*.

3.1. Colocaciones y estrategias

Pongamos por caso la colocación financiera "*tipo de cambio*". Si se recorren las 'pistas' descritas más arriba, es muy probable que los cálculos pasen por alto esta combinación frecuente y típica: el *log likelihood test* no invitará a retener en la lista de palabras clave ni 'tipo' ni 'cambio', no siendo ninguno de los dos típicos de finanzas; como las concordancias se generan a partir de las palabras clave, tampoco se podrán aislar posibles colocaciones frecuentes como "tipo de cambio" de entre las concordancias.
Se puede remediar de dos maneras:

- o bien un experto supervisa la selección de los términos clave y no suprime los dos términos de la lista, conociendo el caso de colocaciones como "tipo de cambio";

- o bien, lo que es mejor, se aíslan primero las colocaciones por temática y se les aplica también el *log likelihood test*.

Otro problema cuantitativo en colocaciones del tipo "*frente a*", requiere aún más la opinión del experto: siendo típica esta combinación tanto del español general como del español económico general y el de finanzas, tampoco será retenida a la hora de aplicar el *log likelihood test*; sin embargo, el experto sabrá que los estudiantes tienden a sustituirla por otros grupos preposicionales menos típicos del área de finanzas, lo cual puede constituir un criterio suplementario para mantener el término en la subtemática correspondiente.

3.2. Introducciones temáticas para "reajustar las cuentas"

Además, *ningún recuento de este tipo es fiable cien por cien*, y las listas de frecuencias tampoco constituyen el evangelio. A fin de anticipar las posibles carencias del cálculo automático, se ha desarrollado otro material didáctico para completar el léxico económico generado exclusivamente a partir del corpus: con una selección del léxico llamado 'básico' según los autores de los manuales y glosarios, se redactaron *introducciones temáticas* sobre cada subtemática (véase, a modo de ejemplo, el cuadro 3). Los términos 'económicos' de las introducciones se presentan con un color diferente –aquí se subrayan-, de manera que el usuario los pueda reconocer como 'económicos'. Se relacionarán mediante un enlace con una ficha terminológica similar a las fichas de los términos provenientes del corpus. En caso de que no se den en nuestro corpus, recibirán la frecuencia '0' en el campo cuantitativo correspondiente.

Introducción temática: La Nueva Economía

1. ¿ Qué es la Nueva Economía?

> La economía cambia continuamente, pero desde hace unos quince años está cambiando muy profundamente. El concepto de 'nueva economía' refiere a un mundo en el que la gente trabaja cada vez más con el cerebro y menos con las manos, es decir, una economía que se basa en el conocimiento y en la información. En ella, el papel de las nuevas tecnologías de información y de comunicación es cada día más importante. En el sector servicios las nuevas tecnologías han llegado a constituir el núcleo de la actividad. La introducción revolucionaria de Internet en el mundo empresarial tuvo como consecuencia la apertura de una nueva vía de contacto con clientes y proveedores. Internet no sólo permite que una infinidad de personas compartan e intercambien información entre ellos, sino que es, además, una plataforma de negocios en línea. El comercio electrónico ofrece grandes posibilidades para las sociedades comerciales. (...)

cuadro 3: extracto de una introducción temática

3.3. Más allá de la obra de referencia

Las introducciones crean también otra oportunidad, a saber: la de convertir la obra de referencia en la parte central de un *entorno de aprendizaje*.

Los términos del léxico que no están en nuestro corpus sino sólo en las introducciones temáticas, reciben el valor "0" para este criterio. Para la base de datos, el usuario dispondrá, aparte del índice alfabético, de otro criterio, a saber: la frecuencia decreciente. Así, para adentrarse en una de las subtemáticas, podrá empezar por la lectura de la *introducción temática* para hacerse una idea global del área; después, podrá decidir él mismo hasta *qué nivel de frecuencia* asimilar los términos del tema correspondiente y, una vez asimilados estos términos, intentar leer alguno de los textos originales sobre el tema en el corpus –o en el periódico del día- y comprobar si estos conocimientos le permiten una comprensión global de los textos periodísticos. En caso de que queden demasiados huecos –que no se deban a una falta de conocimiento del español general-, podrá decidir asimilar más términos de la lista de frecuencias. Así, esta forma de estudio encaja perfectamente en un entorno de aprendizaje autónomo asistido. Por último, se podrá generar una serie de *ejercicios con retroacción* a partir de las concordancias seleccionadas anteriormente.

A fin de crear el entorno de aprendizaje, hace falta construir una plataforma propia, debido a la incapacidad de programas como MultiTerm para combinar base de datos, módulo de ejercicios, corpus virtual de textos, introducciones temáticas, ordenación por frecuencia decreciente y algunos instrumentos de comunicación entre los participantes del proceso de autoaprendizaje asistido.

4. Perspectivas y conclusiones

Cuando termine la financiación del proyecto, esperamos ofrecer en la plataforma de ElektraVoc las fichas terminológicas de las subtemáticas más frecuentes de nuestro corpus económico (dando preferencia, por este orden, a las subtemáticas "finanzas", "empresa" y "economía general"), para que el usuario pueda, de acuerdo con la filosofía global de ElektraVoc, aprender por la práctica y de manera integrada y dinámica en un contexto de aprendizaje autónomo asistido. La obra de referencia temática con los términos económicos más frecuentes de la prensa, por un lado, y de los manuales y glosarios, por el otro, acompañados por contextos auténticos, introducciones temáticas y ejercicios, debería así convertirse en un entorno un de aprendizaje que permita aprender los términos indispensables para comprender la prensa económica. Para conseguir este objetivo, hace falta evitar los problemas antes indicados e integrar las oportunidades que en el camino se presenten.

Bibliografía

Binon, J. & S. Verlinde (1999) La contribution de la lexicographie pédagogique à l'apprentissage et à l'enseignement d'une langue étrangère ou seconde (LES). *Etudes de linguistique appliquée* 116: 453-468.

Buyse, K. & N. Delbecque (2002, i.p.) ElektraVoc. *Aprendizaje del vocabulario español asistido por ordenador*. G. Fabry ed., Jornada de estudios sobre el aprendizaje del vocabulario E/LE. Louvain-la-Neuve, UCL & Embajada de España, http://www.e-montaigne.com.

Buyse, K., P. Muñoz Tavira, E. Roegiest & A. Vandenberghe (2003, i.p.) *Hacia un léxico económico español en soporte electrónico*. Andreu van Hooft & Rob le Pair eds., Foro Hispánico. *El español de los negocios: perspectivas teóricas, investigaciones, nuevas tecnologías y didáctica*. Nijmegen: Universiteit Nijmegen.

Buyse, K. (2003) *Plataforma electrónica del español económico dentro del proyecto ElektraVoc-II*. S. Posteguillo Gómez ed., Internet and Language. Castellón: Universidad de Castellón.

Buyse, K., N. Delbecque & D. Speelman (2003) '*Portavoces'*. Thematische woordenschat Spaans met oefenmodule. CD. Deurne: Wolters-Plantyn.

Corpas Pastor, G. (1992) *Tratamiento de las colocaciones del tipo A+S/S+A en diccionarios bilingües y monolingües (español-inglés)*. Euralex 90. Actas del IV Congreso Internacional, Benalmódena (Málaga), 28 agosto – 1 sept. 1991, Barcelona, Bibliograf, 331-340.