Las lenguas minoritarias europeas y las nuevas tecnologías

Mª Antonia Martí

Profesora titular de la Universitat de Barcelona y directora del CLiC (Centre de Llenguatge i Computació).

1. Lenguas, sociedad y TIC

Las tecnologías de la información y la comunicación, comúnmente conocidas como TIC, están originando cambios en profundidad en el modo en que los humanos nos comunicamos, en todas las modalidades de interacción humana. Si bien el almacenamiento y el acceso a la información ha sido uno de los campos que ha sufrido cambios más radicales en los últimos cinco años, éstos también se están produciendo, en profundidad, en el modo en que adquirimos bienes, contratamos servicios, aprendemos, ejercemos la docencia, nos divertimos y, en suma, interactuamos con nuestro entorno.

Internet es el nuevo marco en el que tiene lugar este nuevo modo de interacción, de manera que la web se ha constituido en el foro de comunicación de la era digital. En él tienen cabida páginas personales, páginas institucionales, blogs, portales temáticos, ofertas de productos, opiniones sobre los mismos, etc., así como aplicaciones para comunicarnos de manera interactiva y directa tanto oralmente como por medio de la escritura. En suma, Internet se ha constituido como un espacio virtual, por tanto, sin limitaciones espacio-temporales, en el que podemos desarrollar una parte importante de las actividades que hasta el momento habíamos venido realizando presencialmente.

Aunque la web es un espacio multimodal en el que conviven imagen, sonido, voz y texto, el modo en que se produce la comunicación, las interacciones y la información es fundamentalmente a través de la lengua hablada o escrita. Los nuevos sistemas de comunicación han dado lugar a nuevas variedades de uso de las lenguas en su forma gráfica. Estamos asistiendo a la aparición de una variante de lengua escrita coloquial, prácticamente inexistente anteriormente, que en aras de la economía de recursos y de la

expresividad combina los caracteres ortográficos con los icónicos y está formulando sus propias leyes de escritura. La existencia de esta variedad de uso, lejos de poner en peligro la calidad de la lengua escrita, puede que se constituya en un criterio nuevo a tener en cuenta para evaluar la vitalidad, el estancamiento o el retroceso de las lenguas.

Desde una perspectiva lingüística, este nuevo escenario plantea una serie de preguntas que nos invitan a la reflexión: ¿Qué lenguas de comunicación están representadas en la web y en qué medida? ¿En qué lenguas están disponibles las aplicaciones que nos permiten interactuar en el entorno digital? ¿Se reproduce en el entorno digital los mismos usos lingüísticos y procesos de sustitución que en las interacciones presenciales? ¿El entorno digital es un nuevo reto que las lenguas minoritarias y minorizadas tendrán que superar, o una oportunidad para su consolidación? ¿Cómo pueden afectar los criterios de rentabilidad al desarrollo de aplicaciones basadas en tecnología lingüística? ¿Qué requisitos debe cumplir una lengua para ser competitiva en el marco de la sociedad digital?

Trataremos de dar elementos para la reflexión sobre estos puntos. No hay respuestas claras, sólo proyecciones de futuro a partir de los datos que actualmente disponemos. El mundo digital es tan reciente y abre perspectivas tan nuevas y complejas que resulta difícil aventurar pronósticos. Lo que si podemos determinar es qué políticas hay que evitar y qué iniciativas hay que promover para el mantenimiento de la pluralidad y riqueza lingüística del planeta. Hay que estar preparado. Empezaremos planteando algunas cuestiones lingüísticas y sociolingüísticas que están en la base del tema que nos ocupa para después ocuparnos de los aspectos tecnológicos.

2. Lenguas minoritarias y lenguas minorizadas

Las lenguas actualmente existentes no comparten el mismo estatus de estabilidad y consolidación. El número de hablantes, su reconocimiento como lengua oficial, el aislamiento o la convivencia con lenguas mayoritarias son factores determinantes para su mantenimiento y supervivencia.

Cuando una lengua no se transmite de padres a hijos, cuando ya no es lengua materna de nadie se considera que ha muerto. En Europa tenemos ejemplos de lenguas que existieron en el pasado y que han desaparecido sin dejar más huella que algunos vocablos en las lenguas actualmente existentes; este es el caso del íbero, del etrusco, del mozárabe o del dálmata, que desapareció a finales del siglo XIX al morir el último hablante que quedaba. Otras lenguas que han desaparecido perviven, de algún modo, en las lenguas a que dieron lugar: este es el caso del latín del que derivan las lenguas románicas (portugués, gallego, catalán, italiano, occitano, español, etc.), el proto-germánico que ha dado lugar a las lenguas germánicas (inglés, holandés, alemán, flamenco, etc.) o el ilirio que ha dado lugar al albanés.

En el mundo existen actualmente alrededor de 6.000 lenguas, la mitad de las cuales no superan los 10.000 hablantes. Se prevé que la mitad de ellas dejarán de transmitirse de padres a hijos a lo largo de este siglo, es decir, desaparecerán. (Crystal, 2001). En Europa se hablan actualmente entre 85 y 90 lenguas [1], pero su estatus es muy diverso dependiendo de si están o no reconocidas como oficiales, de si conviven o no con otras lenguas en un mismo territorio, del número de hablantes y, muy especialmente, del poder político y económico de los estados donde se hablan.

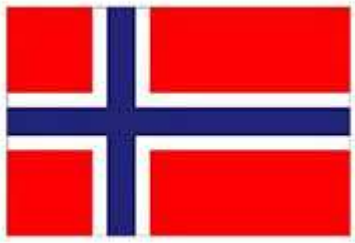
¿Cuál es el futuro de estas lenguas? ¿En qué medida les puede afectar la revolución digital? Para responder a estas preguntas presentaremos primeramente diferentes arquetipos de situaciones que los sociolingüistas han definido para caracterizar el estatus de supervivencia de una lengua. Estos arquetipos se fundamentan en las condiciones sociales, políticas, demográficas y económicas en que se encuentran las lenguas sometidas a estudio. Toda estructura sociolingüística -es decir, toda regulación social del uso lingüístico- se debe analizar en términos de poder y dominación.

En este marco de análisis, se habla de "lengua amenazada" cuando se identifican factores que apuntan hacia su extinción a medio plazo. El factor principal que se tiene en cuenta para considerar que una lengua está amenazada es que los hijos de hablantes nativos estén adquiriendo otra lengua diferente (normalmente predominante en el entorno social y cultural) o dejen de usar su lengua materna en favor de otras lenguas de uso más general, lo que sucede normalmente cuando inician la escolarización y disminuye la influencia del entorno familiar. En general están en esta situación aquellas lenguas habladas por minorías dentro de un estado con una lengua oficial y dominante, situación muy común en el mapa de Europa.

Con el término "lengua minorizada" se hace referencia a un idioma que ha sufrido marginación, persecución o incluso prohibición en algún momento de su historia. Es, por tanto, un concepto que remarca la presencia de una acción coercitiva que lleva a un recorte en su utilización. Este término se diferencia del más ampliamente utilizado "lengua minoritaria", que hace referencia al idioma utilizado por un pequeño número de hablantes. No son, por tanto, sinónimos y no siempre una lengua minorizada es minoritaria, aunque tienda a ello. El proceso de minorización lingüística es el resultado de determinados procesos históricos que han implicado políticas de asimilación y uniformización conscientes y dirigidas. La política lingüística aplicada en el estado español durante el franquismo llevó a un proceso de minorización del resto de lenguas vernáculas que quedaron relegadas al uso familiar; algo parecido ocurrió con la política de rusificación aplicada en el antigua Unión Soviética sobre los estados satélites, en los que se impuso el ruso como materia obligatoria a partir de 1938. La minorización puede darse también en estados democráticos aunque no existan prohibiciones directas, como ocurre en Francia donde en aras de la igualdad no se reconoce oficialmente otra lengua que el francés.

En Europa hay lenguas minoritarias como el noruego (4,5 millones de hablantes) [2], que en ningún caso pueden considerarse como minorizadas puesto que no se ejerce sobre ellas ninguna acción coercitiva. Cabe señalar también que lenguas que estaban en una clara situación de retroceso como el vasco, han incrementado el número de

hablantes y, lo que es más importante, ha aumentado el número de individuos que la tienen como lengua materna. Finalmente, lenguas como el hebreo, una lengua muerta que se había conservado como lengua litúrgica se ha recuperado al declararse como lengua oficial del estado de Israel.



3. Las TIC y las lenguas

Las nuevas tecnologías de la información y la comunicación abren nuevas situaciones de uso lingüístico, dan lugar a nuevas aplicaciones basadas en la lengua y han desarrollado nuevas plataformas de diseminación de la información. Están creando un nuevo escenario, mucho más complejo desde el punto de vista lingüístico, respecto del cual no tenemos todavía la suficiente distancia ni perspectiva como para aventurar cuáles serán sus efectos futuros. Tampoco sabemos cuáles van a ser los efectos que este nuevo escenario va a tener para el mantenimiento de la riqueza lingüística de la humanidad.

Podemos distinguir tres niveles de uso de las nuevas tecnologías: el documental informativo, el de infraestructura básica y el creativo.

3.1. Las TIC como facilitadoras de información

Entendemos por uso "documental informativo" de las TIC las facilidades que las nuevas plataformas de comunicación proporcionan para publicar contenidos y transmitir información. En este sentido se enmarcan iniciativas de carácter colaborativo y participativo como pueden ser las folksonomías o la Wikipedia.

Las folksonomías constituyen un nuevo paradigma de clasificación de la información que permite a los internautas crear libremente etiquetas para categorizar todo tipo de contenidos. Pueden ser utilizadas para difundir contenidos culturales y lingüísticos que de otro modo permanecerían ocultos [3].

La Wikipedia es quizás el exponente más claro de este uso "documental e informativo" de las TIC. Actualmente contiene más de 10 millones de artículos, de los cuales 2,5 millones corresponden a artículos escritos en lengua inglesa y el resto se reparte entre otras 260 lenguas. Constituye una plataforma idónea para el desarrollo de recursos lingüísticos, sin ningún tipo de restricciones como no sean las del propio empeño de sus colaboradores. Hay que tener presente que cada Wikipedia, a parte del valor intrínseco de sus contenidos, constituye un corpus digital de la lengua en que está escrita y, al mismo tiempo, es también un tesauro y una ontología ya que sus artículos están clasificados y organizados jerárquicamente. Como veremos más adelante, estos tipos de

estructura de datos son indispensables para el desarrollo de aplicaciones creativas basadas en tecnología lingüística.

Las estadísticas que periódicamente publica la Wikipedia nos proporcionan elementos para el análisis. En primer lugar, si bien en la Wikipedia están representadas las lenguas mayoritarias, tenemos también ejemplos de lenguas con muy pocos hablantes y claramente minorizadas, como son el aragonés y el veneciano. Es más, el número de artículos no tiene un correlato directo con el número de hablantes. Como podemos observar en la nota al pie de página [4], el chino, la lengua con mayor número de hablantes tiene menos artículos en la Wikipedia que el sueco, el francés o el rumano, que cuentan con un número de hablantes muy inferior.

Por otro lado, el crecimiento de la Wikipedia es muy desigual: lenguas minoritarias y minorizadas pueden presentar un crecimiento mucho mayor que lenguas bien establecidas y con sus derechos plenamente reconocidos. Así, en 2008 el inglés, el alemán, el francés y el polaco habían incrementado el número de artículos en un 1%, mientras que el catalán, el bielorruso, el aragonés y el bávaro habían experimentado un crecimiento del 4% y el vasco y el serbocroata del 5%, el índice más alto.

La presencia de contenidos en Internet puede responder a iniciativas promovidas por parte de organismos oficiales, pero su dinámica es muy abierta y la iniciativa particular y colaborativa por parte de colectivos que pueden estar muy distanciados en el espacio es predominante. Los dos ejemplos que hemos presentado lo ponen de manifiesto. La voluntad de los hablantes y su colaboración desinteresada explican la falta de correlación entre el número de artículos, el crecimiento anual de la Wikipedia y el número de hablantes de las lenguas representadas.

Con todo, estas cifras sólo hacen referencia a las cerca de 270 lenguas representadas en la Wikipedia, menos del 5% de las lenguas existentes. En este 5% están representadas una buena parte de las lenguas que se hablan en Europa (alrededor de 70), tanto minoritarias como minorizadas, pero la simple voluntad y esfuerzo de los hablantes no va a corregir procesos de retroceso y extinción. La presencia en la web es un factor más a tener en cuenta de cara a la caracterización del estatus de una lengua.

Desde una perspectiva más global y no centrada en nuestra realidad inmediata, Koïchiro Matsuura, director general de la Unesco, presentó en 2006 el informe titulado "Hacia las

sociedades del conocimiento", donde se nos alerta sobre el peligro de la extinción de las lenguas y el efecto negativo que tiene el desigual acceso a las nuevas tecnologías. Los datos que proporciona son preocupantes y apuntan que las nuevas tecnologías constituyen un factor que puede acelerar el proceso de desaparición de lenguas. Destacamos algunos puntos:

- Tres de cada cuatro páginas en Internet están escritas en inglés, sin embargo el número de cibernautas cuya lengua materna no es el inglés excede del 50 por ciento, porcentaje que sigue aumentando.
- Sólo un 11% de la población mundial tiene acceso a Internet y el 90% de dichos internautas vive en países industrializados.
- Más de un 90% de las lenguas del mundo (habladas por el 4% de la población del globo) no están representadas en Internet.

Una parte importante de la ciudadanía europea forma parte de este reducido y privilegiado 11% de la población que tiene acceso a Internet, pero Internet nos enfrenta al mundo, y esto nos lleva al predominio de unas pocas lenguas. Si nos atenemos a los datos, casi podría decirse que de una única lengua.

3.2. Infraestructura tecnológica multilingüe

Las TIC han desarrollado programas informáticos de implantación generalizada por parte de todas las personas usuarias con acceso a las nuevas tecnologías. Me refiero con ello a los entornos de edición de textos y a las herramientas de cálculo y de manejo de bases de datos. Inicialmente, las ayudas para facilitar el uso de estos programas sólo existían en versión inglesa; con el tiempo se han realizado versiones a otras muchas lenguas, pero la cobertura lingüística de estas herramientas es insuficiente.

Una de las razones de esta falta de cobertura estriba en que para la traducción de los programas informáticos al mayor número de lenguas posible se requiere de la intervención y colaboración de los sectores público y privado. Esta necesaria conjunción de voluntades deja al margen las lenguas sin estado, las lenguas que están al margen de las nuevas tecnologías y las lenguas para las que no existe una demanda de programas informáticos o la demanda no es suficiente para justificar una inversión desde el sector privado.

Si nuestro análisis se centra en la Comunidad Europea, ¿qué cabe esperar de unos poderes públicos que solo reconocen como oficiales 23 lenguas de los 27 estados miembros? No hay un solo estado europeo que sea totalmente monolingüe, por lo que son muchas las lenguas no oficiales que quedan al margen. A pesar de ello, las empresas privadas han desarrollado versiones de sus programas en muchas de las lenguas europeas que no están oficialmente reconocidas. Puede que la razón sea simplemente de mercado, pero al menos han tomado la iniciativa, a veces por delante de los propios estados e instituciones comunitarias.

Para evitar la "brecha digital", en el informe anteriormente citado, la UNESCO propone varias medidas para evitar el fin de la diversidad lingüística de la que ahora goza el planeta, entre ellas -además de la intervención de los sectores público y privado para la traducción de los programas informáticos- cabe destacar las siguientes: a) la difusión y el uso del software libre y de equipos informáticos asequibles en los países en

desarrollo; b) la promoción de contenidos en Internet en alfabetos diferentes al latino y c) el aprendizaje de dos o tres lenguas desde la educación primaria.

Está por ver el modo y el momento en que se implementarán estas medidas. La "brecha" digital existe. Solo un reducido número de habitantes del planeta tiene acceso a las nuevas tecnologías y, para los que tienen acceso, muchas veces es en una lengua distinta a la suya propia.

3.3. TIC y creatividad

El desarrollo tecnológico ha hecho posible el diseño e implementación de aplicaciones informáticas que tienen como base el tratamiento de lenguaje humano. Estas aplicaciones, que surgen de entornos universitarios, de departamentos de I+D y de pequeñas y grandes empresas tecnológicas, tienen como objetivo facilitar la interacción entre los humanos y las nuevas tecnologías.

Como hemos visto, mientras que la Wikipedia o las folksonomías son el resultado de la acción conjunta y colaborativa de los internautas, las infraestructuras tecnológicas requieren de la conjunción de intereses tanto públicos como privados y la definición de políticas lingüístico-tecnológicas. En el caso de las aplicaciones basadas en procesamiento del lenguaje, sea éste oral o escrito, el factor decisivo para su desarrollo es la iniciativa privada y, por lo tanto, la existencia de un mercado, de una demanda de tales productos. Por lo tanto, su desarrollo va a ser mucho más restrictivo y orientado a aquellas lenguas que disponen de una infraestructura social, económica y política fuerte.

¿En qué consisten tales tecnologías? La más conocida y representativa -y posiblemente la más antigua- es la traducción automática. Los primeros programas de traducción automática se remontan a los años 50, pero no ha habido sistemas realmente operativos hasta los años 90 del pasado siglo. La historia de la traducción automática puede dividirse en dos grandes etapas: en una primera etapa, tenemos los sistemas de traducción basados en el conocimiento; la segunda etapa corresponde a los sistemas de traducción basados en la estadística. Los primeros requieren del desarrollo de gramáticas y léxicos monolingües y bilingües, así como de reglas de transferencia que permiten establecer equivalencias entre las estructuras de las lenguas que se tratan. Los costes de desarrollo son elevados y su eficacia limitada, por lo que suelen restringirse a la traducción de textos de dominios específicos y se han desarrollado para lenguas con

marcado interés económico o político. Los sistemas basados en la estadística requieren disponer de textos bilingües: los programas informáticos infieren las expresiones equivalentes entre una y otra lengua. Estos programas son independientes de la lengua, por lo que su coste de desarrollo es muy inferior y el resultado, razonablemente aceptable. Esta tecnología permite abrir el abanico de lenguas tratadas, ya que los costes de desarrollo son muy inferiores a los de los sistemas basados en el conocimiento, por lo que podrían desarrollarse sistemas de traducción para lenguas minoritarias, minorizadas e incluso para lenguas en claro retroceso.

En el marco de las tecnologías del texto (Badia, 2009), cabe destacar también otras aplicaciones como el resumen automático y la clasificación documental, de gran interés para los centros de documentación, hospitales y, en general para las instituciones que generan diariamente grandes volúmenes de documentación. Los sistemas de extracción y recuperación de información, hasta hace poco de carácter experimental, han empezado a estar disponibles comercialmente [5].

Huelga decir, que estas aplicaciones se han desarrollado inicialmente para la lengua inglesa, que es la que dispone de una infraestructura de recursos de tecnología lingüística más completa y avanzada [6]. Las aplicaciones a otras lenguas suelen hacerse sobre la base de lo que se ha desarrollado para el inglés, por lo que para muchos "tecnología lingüística" y "tecnología de la lengua inglesa" son equivalentes.

Otro campo de aplicación en expansión son las tecnologías de la lengua oral (Llisterri, 2009), cuya base son, por una parte, el reconocimiento del habla, que permite utilizar la voz para muchas de las operaciones que requerirían un teclado y, por otra, la síntesis (o generación) del habla que hace posible la salida vocal.

Estas tecnologías, que son el resultado de décadas de trabajo investigador por parte de lingüistas e informáticos, han empezado a dar sus frutos y existen ya aplicaciones comerciales al alcance de un gran número de personas. La conversión de texto en habla es especialmente útil para las personas con limitaciones de la capacidad visual y para aquellas situaciones en que resulta más práctico escuchar que escribir: el dictado automático de textos, la ejecución vocal de programas informáticos y la realización de acciones como marcar un número de teléfono mediante la voz o buscar un contacto en una agenda electrónica, constituyen claros ejemplos de su interés.

El desarrollo de las tecnologías de la lengua oral y escrita requiere disponer de una infraestructura de recursos lingüísticos potente. Estos recursos son de desarrollo lento y costoso, de modo que difícilmente pueden ser asumidos por parte de la iniciativa privada. Para que una lengua disponga de tales tecnologías han de concurrir diversos factores. Por un lado, la inversión de capital público a través de las universidades y centros de I+D, tanto públicos como privados; por otro, la formación de una masa

crítica de investigadores que permita no sólo aplicar las técnicas desarrolladas para otras lenguas y en otros países, sino también generar nuevo conocimiento e innovar.

4. A modo de conclusión

En este artículo, hemos tratado de presentar un panorama general de las tecnologías de la información y la comunicación y su interrelación con las lenguas. Hemos distinguido diferentes niveles en los que la tecnología está relacionada con las lenguas y hemos visto como el desarrollo de cada nivel depende de factores distintos: de la voluntad de los usuarios, de los poderes públicos o de la iniciativa privada. Como consecuencia, el carácter oficial de una lengua, el hecho de que tenga o no un estado que la haya adoptado como propia, es decisivo para el desarrollo de una política tecnológica que permita disponer de aplicaciones y entornos de usuario para esta lengua.

Aunque muchas de las lenguas europeas disponen de aplicaciones tecnológicas, la tecnología lingüística ha surgido y se ha desarrollado fundamentalmente para el inglés. Si no se aplican políticas correctivas, sólo un número muy reducido de lenguas (probablemente no más del 1% de las actualmente existentes) desarrollarán este tipo de aplicaciones.

El proceso de normalización lingüística tiene como finalidad que una lengua sea usada en toda su variedad de usos y registros. Cabe preguntarse si en la actualidad no se está produciendo una situación de "diglosia digital" a nivel global: las lenguas maternas se utilizan para la comunicación directa y recurrimos al inglés para la comunicación digital.

5. Referencias

- Badia, T. (2009) "L'impacte social de les tecnologies de la llengua", a M. A. Martí (ed.) *Llengua, Societat i Comunicació* num 7. www.ub.edu/
- Crystal D. (2001). *La muerte de las lenguas*. Cambridge University Press, Madrid.
- Llisterri, Joaquim (2009) "Les tecnologies de la parla", a M. A. Martí (ed.) *Llengua, Societat i Comunicació* num 7. www.ub.edu/
- Martí (ed.) (2003) *Tecnologías del lenguaje*. Editorial UOC, Barcelona.
- Matsuura, Koïchiro (2006) "Hacia las sociedades del conocimiento", ocu.uni.edu.pe.

6. Páginas de interés:

- www.worldlanguage.com/
- www.Ethnologue.com
- es.wikipedia.org/wiki/Meta

Notas:
[1] Solo de la familia indoeuropea, la más numerosa en nuestro continente tanto en número de lenguas como de hablantes, existen ente 50 y 55 lenguas. Otras 30 aproximadamente forman parte de las familias caucásica, urálica, altaica y afroasiática. El vasco es una lengua aislada, ya que todavía no ha podido establecerse su filiación.

[2] El concepto de lengua minoritaria es relativo. El noruego es una lengua minoritaria comparada con el inglés, el chino e incluso el francés o el alemán. No lo será si la comparamos con el luxemburgués (300.000 hablantes), el escocés (100.000 hablantes) o el friulano (600.000).

[3] Una de las más conocidas es http://delicious.com, un portal de contenidos culturales con una importante presencia de recursos lingüísticos digitalizados.

[4] Las 25 wikipedias con mayor número de artículos a 30 de mayo de 2008 eran: inglesa (2.393.501), alemana (755.715), francesa (664.938), polaca (505.689), japonesa (494.117), italiana (458.487), holandesa (Nederlands) (441.369), portuguesa (381.365), española (365.560), rusa (286.824), sueca (284.066), china (178.134), noruega (168.045), finesa (164.000), catalana (117.310), volapük (116.292), ucraniana (111.784), rumana (108.772), turca (108.132), esperanto (99.215), checa (98.188), eslovaca (96.321), húngara (95.879), danesa (86.905), indonesia (82.886).

[5] Véase A. Martí (ed.) *Las tecnologías de la lengua*, para una visión general de las aplicaciones de tecnología lingüística.

[6] Son recursos de tecnología lingüística los léxicos y gramáticas computacionales, las ontologías y los tesauros digitales, los corpus anotados, los diccionarios electrónicos, etc.