Una aproximación a la dificultad de los textos. El proyecto CÓDIGO

José Jorge Amigo Extremera
Grupo de investigación PETRA
Universidad de Las Palmas de Gran Canaria
josejorge.amigo@cogtrans.net

Resumen

Exposición de las principales características del proyecto trianual «Caracterización Objetiva de la Dificultad General de los Originales» (CÓDIGO) subvencionado por el Plan Nacional I+D (MiCINN FFI2010-15724).[1] El proyecto cuenta con veinticinco investigadores de las universidades de Las Palmas de Gran Canaria, Autónoma de Barcelona, Granada, Vigo, País Vasco y Leipzig.

Tras presentar los fundamentos teóricos de este proyecto (fórmulas de legibilidad y procesos cognitivos de lectura y comprensión), se describe la metodología (banco textual y pruebas a informantes). Finalmente, se esbozan algunos resultados provisionales.

En el plano científico, el proyecto busca aprehender mejor los procesos de comprensión al traducir. En el plano técnico, pretende desarrollar una aplicación informática de análisis de textos o fragmentos textuales originales y traducciones en español, francés e inglés para predecir su dificultad. En el plano profesional, CÓDIGO se propone contribuir a los criterios en torno a dos parámetros: *a)* cálculo de tarifas en las traducciones y *b)* determinación del perfil idóneo del traductor (*pericia*) al que encomendar un texto.

Palabras clave: dificultad, textos originales, mercado profesional

1. Introducción

El proyecto trianual «Caracterización Objetiva de la Dificultad General de los Originales» (CÓDIGO) tiene como principal objetivo desarrollar una aplicación informática de análisis de textos (o fragmentos textuales) y traducciones en inglés, español y francés para predecir su dificultad. No se descarta incluir otros en el futuro, siempre en la variedad lingüística de mayor volumen editorial, para mejorar su fiabilidad. La aplicación informática resultante tendrá formato web, y será de acceso libre para traductores autónomos y

[1] Este artículo resume unos contenidos y amplía otros del proyecto CÓDIGO, recogidos en la memoria técnica de Ricardo Muñoz Martín para la convocatoria 2010 de ayudas para Proyectos de Investigación Fundamental no orientada del Ministerio de Ciencia e Innovación.

empresas dedicadas a servicios de traducción y terminología, para incidir en la calidad de las traducciones y en la economía del sector (§ 2, Muñoz 2010:2).La información se ofrecerá al usuario de dos formas complementarias: *a)*tres índices autónomos de dificultad (léxica, sintáctica y textual) en una escala de 0 a 100, derivados del análisis cuantitativo de valores textuales y valores de pruebas de velocidad de lectura y de comprensión a 72 informantes y *b)* una lista de 150 textos de un corpus de creación propia por orden de dificultad decreciente, entre los que se ubicará el texto.

2. Objetivos e hipótesis

CÓDIGO pretende contribuir a resolver dos lagunas en la metodología de la investigación empírica en traducción: *a)* criterios objetivos para determinar la dificultad de los textos utilizados en los experimentos y *b)* criterios homogéneos para seleccionar y caracterizar sujetos (Muñoz 2009). Asimismo, busca un conocimiento replicable de los procesos de comprensión al traducir con un método poco explorado que amplía y puede mejorar las tradicionales fórmulas de legibilidad. Este objetivo general tiene una aplicación paralela en el mercado de la traducción profesional, que adolece de dos problemas crónicos: (1) en el cálculo de tarifas de las empresas de traducción apenas considera la dificultad real de los textos, diluida en vaguedades como *textos generales*, *especializados* y *semiespecializados* y (2) en los clientes legos, valorar intuitivamente la dificultad de los originales desemboca a menudo en una infravaloración de necesidades, lo que puede desembocar en traducciones de baja calidad. Una aplicación abierta de orientación en las necesidades podría ayudar a separar el mercado profesional de la amplia y creciente franja de traducción *amateur,* aunque la dificultad de un texto no es, desde luego, el único factor que incide en ello. Más bien, es uno de los factores importantes que suele estar ausente de los cálculos.

Los factores habituales en el cálculo de tarifas son heterogéneos además. A las confusas y borrosas categorías sobre la dificultad de los textos (1) se suelen sumar longitud, formato, tema, difusión y combinación lingüística; a veces, otros. En (2) se da cuenta de un conflicto que CÓDIGO pretende solucionar: al minusvalorar la dificultad de los textos, algunos usuarios (o empresas) encargan tareas de traducción a personas —profesionales o no— cuya pericia aún no alcanza para ofrecer resultados satisfactorios. Incluso asumiendo que un buen trabajo de documentación podría mejorar el resultado, tampoco se da una coherencia pareja con la negociación de plazos de entrega:

> Algunos clientes han vuelto con nosotros a los dos o tres meses para actualizar alguna documentación que han traducido otros, y ves que la traducción es penosa. Incluso en algún caso reconocen que ese precio más barato les resultó mucho más caro. Muchas veces, hacen caso a un compañero de trabajo que dice que conoce a un traductor fenomenal (normalmente, algún familiar o amigo) que termina siendo un fiasco.
>
> (Entrevista de Alberto Caballero a Juan José Arevalillo, 2002:18)

CÓDIGO plantea cuatro hipótesis cuya validez se comprobará por medio de cuatro objetivos: *a)* compilar un banco textual para cada lengua del proyecto, *b)* crear una base de datos léxicos para cada lengua del proyecto, *c)* realizar pruebas de caracterización

de informantes y de lectura, comprensión y traducción de varios fragmentos del banco textual y *d)* desarrollar una aplicación de análisis textual en Internet partiendo de los resultados de la base de datos léxicos y de las pruebas a informantes.

Actualmente se asume que el traductor es capaz de documentarse a fondo sobre los temas que traduce y que es un excelente lector no especialista. CÓDIGO comprobará si esto es realmente así y si existe un desarrollo paralelo con su perfil profesional. Para ello, toma una muestra de estudiantes novatos de traducción, estudiantes avanzados y traductores con al menos cuatro años de experiencia profesional. Así, la primera hipótesis del proyecto es la siguiente: *los profesionales mostrarán mejores destrezas y comprensión que los estudiantes novatos*.

Características de los informantes como la capacidad de memoria de trabajo, la velocidad de procesamiento mental y la comprensión verbal en lengua materna se miden con apartados específicos del test WAIS-III (y no el WAIS-IV porque aún no está normalizado en muchas lenguas).Como existen antecedentes que apuntan a que los resultados pueden ser positivos en aspectos tales como diferencias de memoria en intérpretes, traductores y bilingües, nos concentraremos en otro punto más oscuro, un fenómeno de reciente interés: las creencias implícitas que guían la conducta. La segunda hipótesis es que *habrá una correlación entre las creencias implícitas de los sujetos y su rendimiento al leer y al traducir*. Para probar su verosimilitud, CÓDIGO intenta tipificar a los sujetos por sus creencias y explorar las relaciones con su rendimiento en pruebas de velocidad de lectura, comprensión y traducción.

La tercera hipótesis es que *los diferentes volúmenes relativos de las palabras relacionadas por su significado (sinonimia, antonimia, hiponimia, hiperonimia, ideas afines) permitirán predecir parte de la variación en la dificultad*. En este punto, se proponen fórmulas separadas para los aspectos léxico, sintáctico y suprasintáctico (textual). Así, por ejemplo, se comparan parámetros tales como la complejidad de la palabra (si es deíctica, abstracta o comodín) con la frecuencia.

Por último, frente al carácter monolingüe de la mayoría de las fórmulas de legibilidad, CÓDIGO supone un esfuerzo plurilingüe con índices para el español, el francés y el inglés. Se parte de la hipótesis de que *la dificultad de los documentos comparables en varias lenguas guarda correlación*. De ahí, se asume que los documentos son comparables por su uso: los textos que se utilizan en las mismas circunstancias tienden a tener un nivel similar de dificultad. Para ello, en este proyecto se ha compilado un banco de traducciones para comparar sus índices y también los de los originales de la lengua respectiva.

3. Fundamentos teóricos

El proyecto *a)* está dirigido a adultos que trabajan en el entorno de las industrias de la lengua (distintos de la docencia) y no a niños, como la mayoría de las fórmulas de legibilidad y *b)* introduce parámetros de cohesión y de dificultad del procesamiento mental (léxicos, sintácticos y suprasintácticos). Estas características confieren al proyecto CÓDIGO un sustento teórico sólido y replicable empíricamente.

3.1. **Destinatarios**

Tradicionalmente, las fórmulas de legibilidad han evaluado la capacidad de lectura de sujetos en el ámbito educativo (niños, estudiantes de secundaria, aprendices de segundas lenguas) y menos en personas mayores (los trabajos en el entorno militar son bastante significativos, Dubay 2004). Dubay (2004: 54) ofrece una muestra detallada de las áreas de trabajo más conocidas:

> political literature (Zingman 1977), corporate annual reports (Courtis 1987), customer service manuals (Squires and Ross 1990), drivers' manuals (Stahl and Henk 1995), dental health information (Alexander 2000), palliative-care information (Payne et al. 2000), research consent forms (Hochhauser 2002; Mathew 2002; Paasche-Orlow et al. 2003), informed consent forms (Williams et al. 2003), online health information (Oermann and Wilson 2000), lead-poison brochures (Endres et al. 2002), online privacy notices (Graber et al. 2002) medical journals (Weeks and Wallace 2002), environmental health information (Harvey and Fleming 2003) and mental-health information (King et al. 2003).

Muñoz & Castro (2012: 501) definen los objetivos de las fórmulas de legibilidad y el *modus operandi* tradicional en las investigaciones en las que se aplican:

> El objetivo de estas fórmulas es determinar la dificultad de los textos, casi siempre con el propósito de *a)* ofrecerlos como lectura o como examen en uno u otro estadio educativo, *b)* mejorar su legibilidad (reescribiéndolo cuando no se considera adecuado), o *c)* determinar con ellos la capacidad de comprensión lectora de los sujetos. Estas fórmulas suelen escoger parámetros textuales computables, como la longitud de las palabras, que se utilizan para caracterizar un número variable de textos. Después, un grupo de informantes lee esos textos para ejecutar alguna de estas tareas: *1)* señalar intuitivamente su grado de dificultad en una escala, *2)* responder a un cuestionario de preguntas de comprensión, o bien *3)* rellenar los huecos de textos tipo *cloze* con la misma información que los respectivos originales. Con los resultados de este segundo paso, los investigadores proceden entonces a buscar una fórmula matemática que prediga la clasificación de los sujetos a partir de los resultados de la primera fase. Finalmente, la fórmula se aplica a nuevos textos, para predecir su grado de dificultad.

Estudiar la legibilidad de los traductores tiene su justificación económica, puesto que el porcentaje de traducciones se ha mantenido desde 1990 en torno al 25% (un 26% en el año 1990, un 25 % en 2004 (Pegenaute 2004: 591) y un 26,8 % en el año 2007) de la producción editorial, dato que refleja las dimensiones de la inversión editorial en libros traducidos y la consolidación de la demanda por parte del lector de libros traducidos (Ministerio de Cultura 2008: 8).

3.2. **Parámetros de cohesión y dificultad del procesamiento mental**

Desde una perspectiva cognitiva, la frecuencia del léxico es una variable relevante para determinar la legibilidad (Muñoz 2010: 3). En este proceso entran en juego la velocidad de lectura, la eficiencia de la actividad cerebral y la precisión en el recuerdo (Douglas 2009). Otros parámetros léxicos introducidos por CÓDIGO son la correlación entre palabras *vacías* y *llenas* (Bradac, Davies & Courtright 1977) y el volumen de *palabras abstractas* (Flesch 1948) y *vagas* (Lewis 2006).

Para Kintsch (1988), el proceso de lectura es el resultado de la interacción de tres vertientes: *a)* reconocimiento y decodificación de las unidades lingüísticas; *b)* interrelaciones entre proposiciones (construcción de la representación mental del texto completo: *textbase*, Kintsch 1988:164) y *c)* representación completa de la situación que presenta el texto (*situation model*, Kintsch 1988:180). Especialmente durante las dos primeras fases, el lector construye una base de conocimiento relevante para comprender un texto, proceso similar a cómo se recuerdan listas de palabras (*generation-recognition principle*, Kintsch 1988:179). CÓDIGO complementa esta tercera vertiente con información demográfica y sociolingüística, el análisis de las creencias implícitas y varios índices psicométricos del test WAIS III.

La cohesión textual también es un parámetro de gran relevancia para los estudios cognitivos de la comprensión (Muñoz 2010: 4). Cabe destacar aquí las aportaciones del *análisis semántico latente* (LSA, siglas en inglés, Folz et al 1998, Landauer et al 1998) y Coh-Metrix (McNamara et al 2002, Graesser et al 2004, Crossley et al 2007). El LSA es un método utilizado para extraer y representar el significado derivado del uso de una palabra en función de los contextos en que aparece por medio de un complejo análisis estadístico que se aplica a corpus de grandes dimensiones. Su estrategia para determinar la dificultad de un texto consiste en crear un amplio corpus de textos paralelos y calcular la coocurrencia de cada palabra en cada uno de ellos. Su principal problema es que no contempla la dificultad como función de la atipicidad de un texto (menor coocurrencia de palabras), sino que trata los textos como elementos aislados. De su aproximación, CÓDIGO recoge la estructura de la matriz de datos y las pruebas de comprensión con procedimiento *cloze* (Muñoz 2010: 4).

Coh-Metrix es una iniciativa para estudiar la cohesión desarrollada por el Departamento de Psicología de la Universidad de Memphis[2]. En breve, trata de comprobar varios parámetros por medio de pruebas de comprensión (preguntas verdadero/falso) y abundantes pruebas de lectura en estudiantes. Pese a sus ventajas, las críticas aluden a la complejidad al mostrar los resultados y la redundancia de muchos de sus índices (Swenson 2008:37). CÓDIGO reduce los parámetros e intenta mostrar los resultados de forma clara y comprensible para el usuario lego. Asimismo, se explora también la *correfencialidad,* definida como «la relación que establece un nombre, pronombre o sintagma nominal con una proposición del mismo texto que remite a una misma entidad» (Muñoz 2010: 4).

4. Materiales y métodos

La metodología utilizada en el proyecto CÓDIGO obedece a los objetivos expuestos en § 2.

[2] <http://cohmetrix.memphis.edu/cohmetrixpr/index.html> (Última consulta, 21 de enero de 2013)

4.1. **Compilar un banco textual para cada lengua del proyecto**

Actualmente, el grupo del proyecto está caracterizando un banco textual (o *corpus*) de 100 textos por lengua para ofrecer sus valores de dificultad en la aplicación informática final. Estos valores de dificultad se basan en el análisis de pruebas (de lengua materna, lengua extranjera y traducción) de 40 fragmentos textuales en 72 sujetos. Dentro de cada banco textual, para este proyecto se ha establecido una tipología que contempla textos *canónicos* y *traducidos*. A efectos de nuestra base de datos, se entiende por *texto canónico* todo original prototípico de gran uso susceptible de traducción. Dentro de los textos canónicos hay un subgrupo de *textos experimentales* que sirve de base para las pruebas. Los *textos traducidos* tienen traducciones publicadas a todas las lenguas del proyecto; en este caso, el criterio es la frecuencia de traducción. La tabla 1 detalla la tipología textual de CÓDIGO.

				palabras
canónicos	*experimentales*	1.	Manual de bachillerato	500
		2.	Artículo de opinión sobre política nacional	600
		3.	Conjunto de siete noticias breves	700
		4.	Prospecto farmacéutico	800
		5.	Folleto turístico	900
		6.	Formulario	1000
		7.	Contrato	2000
		8.	Manual de instrucciones	3000
		9.	Artículo de investigación empírica	4000
		10.	Narración literaria breve	5000
	otros	90 textos sugeridos por empresas, asociaciones profesionales y traductores autónomos. El número máximo de textos literarios (o secciones completas de ellos) es 10.		
traducidos	25 textos originales y 50 traducciones de las demás lenguas.			

Tabla 1. *Tipología textual de CÓDIGO*

La estrategia para seleccionar los *textos experimentales* sigue la del proyecto CroCo en cuanto a la representatividad textual (Neumann & Hansen-Schirra, 2005:s.p.).

> While we can, for instance, count all people living on a given stretch of earth, we cannot count all texts produced within a given period of time (if we do not want to narrow the sample down to a restricted author or author's collective). One might think, merely increasing the size of the resource as much as possible both in terms of text types covered and of number of words contained may ultimately equal representativeness. We content that a smaller corpus which is well-designed and annotated is preferable to a large one which may contain material not adding any information to the research question.

Así, desde este proyecto asumimos que no es posible reunir todos los textos en un corpus, pero sí se puede extraer una muestra representativa de los más comunes.

4.2. **Crear una base de datos léxicos para cada lengua del proyecto**

La base de datos léxicos se compone de términos e informaciones sobre ellos, y es la base del *categorizador* (§ 5) y lo será del analizador textual resultante. Los términos pertenecen a corpus de referencia (CREA, CERF y ANC) y se cotejan de acuerdo con los siguientes diccionarios en formato electrónico:

	corpus	**diccionarios**
español	CREA (RAE)	*DRAE, Moliner, Larousse*
francés	CERF	*Le Petit Robert, TLFI, Larousse*
inglés	ANC	*Merriam-Webster's, New Roget's Thesaurus*

Tabla 2. *Fuentes de referencia lexicográfica en CÓDIGO*

Las bases de datos léxicos estructuran las palabras en tres categorías: *a)* categorías cerradas invariables, *b)* categorías cerradas variables (pronombres y determinantes) y *c)* categorías abiertas o léxicas.

Las **categorías cerradas invariables** constituyen listas exhaustivas de ocurrencias y lemas. Añadir los lemas es una estrategia importante de CÓDIGO, ya que permiten establecer relaciones semánticas entre términos. Dentro de esta categoría se incluyen adverbios (lugar, tiempo, modo y cantidad), conectores (copulativos, adversativos, disyuntivos, de relativo, etc.), cuantificadores (cardinales y ordinales) e interjecciones. Estas listas distinguen por género, número y caso; y también contemplan palabras vagas o comodín (*cosa*, *asunto*).

Las **categorías cerradas variables** incluyen pronombres y determinantes etiquetados convenientemente para procesarlos según las siguientes categorías: definidos, indefinidos, sujeto, objeto con preposición, objeto sin preposición, posesivos, demostrativos e interrogativos.

Las **categorías abiertas o léxicas** se refieren a los 10000 lemas más frecuentes de cada lengua en los corpus de referencia, marcados en tramos de 500, y la lista exhaustiva de sus formas flexivas, sin derivación. Incluyen adjetivos (comparativos, superlativos, simples, compuestos, relacionales), nombres comunes y verbos. A estas categorías se añaden listas separadas para nombres propios y apellidos, topónimos, cognados y falsos amigos.

4.3. **Realizar pruebas de caracterización de informantes y de lectura, comprensión y traducción de varios fragmentos del banco textual**

Los informantes se dividen en tres grupos: profesionales, estudiantes novatos (por debajo del cuarto semestre de formación universitaria en traducción) y estudiantes avanzados (al menos en su séptimo semestre). Los informantes realizarán las pruebas desde su domicilio por medio de una aplicación web. En CÓDIGO se han diseñado seis pruebas comunes para expertos y estudiantes:

(1) Velocidad de lectura (lengua A y lengua B)
(2) Comprensión lectora (lengua A y lengua B)

(3) Traducción
(4) Índices psicométricos
(5) Creencias implícitas
(6) Cuestionario para perfil demográfico, laboral y sociolingüístico

Los estudiantes realizan además una prueba estándar que mide las competencias lingüísticas en lengua extranjera. Las tres primeras se realizan sobre segmentos textuales completos de entre 500 y 700 palabras extraídos de los textos canónicos. Las pruebas se realizan sobre los textos experimentales de la lengua materna (A) y la mitad de los de la lengua extranjera de cada sujeto (B).

(1) Velocidad de lectura. Comprende cuatro pruebas complementarias: *lectura inmotivada* (3 minutos, el informante ha de señalar la última palabra que ha leído), *lectura autoadministrada por párrafos* (3 minutos, el programa computa la duración de lectura de cada párrafo, que aparece en pantalla cada vez que el sujeto pulsa la barra espaciadora), *lectura autoadministrada por oraciones* (3 minutos, similar a la anterior, solo que la interfaz muestra el texto por oraciones) y *lectura motivada* (3 minutos, tras la lectura, también autoadministrada por párrafos, el sujeto deberá responder a unas preguntas de comprensión).

(2) Comprensión lectora. Estas pruebas se llevan a cabo tras leer el texto o durante su lectura, tras un mínimo de 250 palabras. Comprende tres pruebas: *text cloze* (10 minutos.) Tras la lectura, se reproducen unas 250 palabras con huecos en aproximadamente una de cada cinco *palabras llenas*, que el informante debe cubrir, *verificación oracional* (9 minutos). Tras un cambio de pantalla, el informante tiene que responder a una serie de preguntas en las que se introducen variantes oracionales de un mismo texto: originales, parafraseadas, erróneas y distractoras.) Esta prueba sigue la metodología expuesta en Royer, Green & Sinatra (1987) y emplea una técnica que mide la comprensión de un texto por medio de cuatro variaciones oracionales:

> *Originals* are exact copies of passage sentences. *Paraphrases* have the same meaning but most of the words are changed. A *meaning change* item contains many of the same words as an original sentence but means something different. A *distractor* test item concerns the same topic but differs in meaning and wording from any passage sentence.
> (Royer, Green & Sinatra 1987:415, cursivas mías)

Para ilustrar cómo se administra esta prueba, tomaremos como ejemplo un fragmento de una noticia publicada en la versión digital del diario *El País* el 11 de mayo de 2009[3]. El texto se titula *Arte contra el fracaso escolar* y no pertenece al banco textual, tan solo se usa como ilustración para este artículo:

> **Fragmento textual original**
> Un alumno líder de un aula, extrovertido y carismático, pero a la par alborotador y mal estudiante, puede empezar a obtener estupendos resultados tras ser elegido por sus compa-

[3] <http://elpais.com/diario/2009/05/11/educacion/1241992802_850215.html> (Última consulta, 21 de enero de 2013)

> ñeros para protagonizar una obra de teatro. Es una experiencia real vivida por los expertos que ahora han puesto en marcha un programa —entre la Fundación Hogar del Empleado (Fuhem), la Organización de Estados Iberoamericanos (OEI) y la Fundación Giner de los Ríos— para concienciar y orientar a los docentes sobre el papel del arte a la hora de sanar los malos expedientes y de educar a ciudadanos responsables. En su programa, de formación y jornadas, han contado experiencias como la protagonizada por un grupo de alumnos de Parla, localidad de 107.000 vecinos al sur de Madrid. Nunca habían cogido un avión y apenas habían salido de su ciudad.
>
> (Silió 2009: en línea)

Este fragmento textual puede manipularse del siguiente modo para la prueba de verificación oracional:

> **Fragmento textual manipulado**
>
> El mero hecho de que unos alumnos elijan como protagonista de una obra de teatro al líder de la clase, el gamberro y mal estudiante, pero sociable y carismático, puede repercutir muy favorablemente en su rendimiento **(OP)**. Es una experiencia real vivida por los expertos que ahora han puesto en marcha un programa —entre la Fundación Hogar del Empleado (Fuhem), la Organización de Estados Iberoamericanos (OEI) y la Fundación Giner de los Ríos— para concienciar y orientar a los docentes sobre el papel del arte a la hora de sanar los malos expedientes y de educar a ciudadanos responsables **(OO)**. En su programa de formación y jornadas, han contado experiencias como la protagonizada por un grupo de alumnos de Alcorcón, localidad de 170.000 vecinos al suroeste de Madrid **(OE)**. Se mostraron muy ansiosos e inseguros durante los ensayos de la obra que habían elegido para ellos **(OD)**.
>
> oración original **(OO)**, oración parafraseada **(OP)**, oración errónea **(OE)**, oración distractora **(OD)**

El texto manipulado se presenta oración por oración y los informantes deben decidir si esa oración estaba en el texto. Se aceptan la original y la paráfrasis.

Finalmente, los sujetos rellenan un cuestionario de opinión (13 minutos) sobre la dificultad de los textos leídos y elaboran una escala de dificultad propia. Los resultados de estas pruebas se distribuyen en cuatro índices separados: índice objetivo de comprensión *cloze*, verificación oracional, índices subjetivos inmediatos de comprensión y dificultad e índices subjetivos retrospectivos de dificultad y comprensión.

(3) Traducción. Esta prueba se lleva a cabo con la mitad de los textos de la lengua extranjera del informante no incluidos previamente en las pruebas de lectura para ese sujeto. Se presenta una sección con los dos párrafos iniciales de un texto y se debe traducir el segundo párrafo sin consultar ninguna fuente. Aquí se mide el tiempo empleado (índice de velocidad), la calidad (evaluada por tres expertos en corrección ciega), el tiempo aproximado de lectura al traducir y la dificultad de los párrafos.

(4) Índices psicométricos. Estos índices se obtienen de nueve pruebas del test WAIS III. Se obtienen como resultado tres índices que caracterizan a cada informante dentro del grupo y frente a la población general, de especial interés para la traductología (Bolaños 2012: 473-474): *a)* comprensión verbal (tests de semejanzas, información y comprensión), *b)* memoria de trabajo (tests de retención de dígitos, aritmética y series de letras y

números) y *c)* velocidad de procesamiento (tests de búsqueda de símbolos y codificación de dígitos y símbolos).

Para alejarlo de su uso como prueba de inteligencia, se transgreden algunas normas de aplicación. En concreto, se administra visualmente, *online* y en diversas sesiones de 10 minutos entre otras pruebas y en un orden distinto al del WAIS-III. No obstante, se hará una prueba de re-test con el método tradicional a parte de los informantes para posibilitar un contraste con los índices generales de referencia.

(5) Creencias implícitas. Siguiendo la línea iniciada por Presas y Martín de León (2011), se trata de completar oraciones truncadas que se corresponden con afirmaciones sobre creencias relativas a los conceptos de *lengua*, *lenguaje*, *significado*, *comunicación* y *cultura*. Se desglosa en dos pruebas: *a)* conceptos básicos (6 minutos) y *b)* afinidad ideacional (12 minutos). La prueba de conceptos básicos presenta oraciones de tres tipos:

1. Declaraciones en lenguaje formal: *«lo más importante de la traducción es _____»*
2. Declaraciones en lenguaje coloquial: *«para mí, lo más fácil de traducir es _____»*
3. Comparaciones: *«Traducir se puede comparar con _____»*

La prueba de afinidad ideacional consiste en escoger las 6 afirmaciones que más se ajustan a las creencias del informante de un grupo de 12. Estas declaraciones son de tres tipos:

1. Afirmaciones literales sobre cómo enfrentarse a la tarea de traducir: *«Hay que traducir el sentido»*
2. Expresiones metafóricas que suponen teorías de traducción implícitas: *«Traducir es negociar entre culturas»*
3. Principios teóricos asociados a teorías de la traducción: *«Traducir es una práctica deliberada»*

(6) Cuestionario demográfico, sociolingüístico y laboral. Este cuestionario se divide en dos partes: uno inicial que dura 10 minutos y en el que se recaba información demográfica, sociolingüística y laboral y otro final que dura 25 minutos y en el que se solicita información cualitativa sobre la profesión, el estado de la lengua materna, las razones para estudiar o ejercer la traducción, opinión personal, etc.

4.4. **Desarrollar una aplicación de análisis textual en Internet partiendo de los resultados de la base de datos léxicos y de las pruebas a informantes**

El desarrollo de esta aplicación constituye la fase final del proyecto CÓDIGO y consta de tres módulos: el de etiquetado y lematización, el de análisis textual propiamente dicho y el de cómputos estadísticos, que muestra los resultados de los dos primeros.

El módulo de análisis textual calcula tres índices (en número y porcentaje): léxico, sintáctico y textual. El índice léxico considera, de entrada, los valores absolutos de los siguientes parámetros de cada texto: ocurrencias, lemas, palabras vacías, palabras llenas, ocurrencias y lemas únicos de palabras llenas presentes en la base léxica (que no se repiten), ocurrencias ausentes de la base, ocurrencias únicas ausentes de la bases, categorías y subcategorías incluidas en la base léxica (§ 4.2.) y el valor de ambigüedad de las palabras llenas. Asimismo, calcula la frecuencia de vocabulario para cada ocurrencia,

asignándole el valor de su lema. El índice sintáctico calcula el número de proposiciones, cláusulas, períodos y oraciones (incluidos fragmentos y títulos). El índice textual computa el número de párrafos y epígrafes, así como su longitud en oraciones, periodos, claúsulas, proposiciones, palabras llenas y ocurrencias.

5. Relación de resultados provisionales

A fecha de enero de 2013, el equipo de CÓDIGO cuenta con una serie de resultados provisionales que nos van a permitir realizar las pruebas a sujetos próximamente.

El **banco textual** para cada una de las lenguas está terminado y comprende las categorías textuales expuestas en § 4.1. La labor de documentación no ha sido tarea desdeñable: solicitudes de trabajo y formularios de inmigración y visados de España, Francia y Estados Unidos, textos muy traducidos. Por ejemplo, las noticias breves (similares a la de § 4.3.) debían satisfacer la condición de no propiciar claramente posicionamientos ideológicos que pudieran incidir en la conducta, por lo que se ha recurrido a muchos textos expositivos (por ejemplo, recomendaciones sobre qué visitar en un país extraídas de artículos editoriales). Los textos han requerido una extensa labor manual de adecuación y cambios de formato, ya que el *analizador textual* solamente admitirá documentos en formato .txt. Además, se han hecho correcciones ortográficas, suprimido ilustraciones y gráficas y codificado las palabras destacadas tipográficamente y los textos insertos, paratextos (*abstracts*, entradillas), notas a pie y bibliografías.

Las **bases de datos léxicos** se han elaborado seleccionando y combinando lemas de los corpus de referencia (§ 4.2), según criterios de frecuencia y dispersión del uso. Los grupos lingüísticos de cada lengua (uno por lengua, de cinco investigadores), ha elaborado listas de lemas, adaptándose a sus características sintácticas y morfológicas y teniendo en cuenta la propuesta de categorización general para las tres lenguas desarrollada para CÓDIGO (adaptada de Muñoz 2010):

1. Adjetivo de grado
2. Adjetivos calificativos
3. Adjetivos determinativos
4. Otros adjetivos
5. Adverbios
6. Adverbios interrogativos
7. Adverbios oracionales
8. Adverbios relativos
9. Otros adverbios (adverbio extranjero)
10. Afijos
11. Artículos
12. Conjunciones coordinantes
13. Conjunciones subordinantes
14. Contracciones
15. Fraseología
16. Locuciones adjetivas

17. Locuciones adverbiales
18. Locuciones conjuntivas coordinantes
19. Locuciones conjuntivas subordinantes
20. Locuciones extranjeras
21. Locuciones onomatopéyicas e interjectivas
22. Locuciones prepositivas
23. Locuciones sustantivas
24. Nombres comunes
25. Nombres propios
26. Numerales
27. Preposiciones
28. Pronombres
29. Pronombres personales
30. Segmentos no léxicos
31. Verbos

En los casos pertinentes, se han recogido también las variantes de número y género. A la base principal se añaden las listas de palabras específicas, tales como las siguientes, de las que actualmente dispone ya el grupo de lengua inglesa:

- 10000 lemas más frecuentes en lengua inglesa.
- 5000 nombres y apellidos más frecuentes.
- Más de 300 *falsos amigos* que conciernen a todas las lenguas del proyecto (por ejemplo, *accuse*, *demand*, *remove*).
- Combinaciones más frecuentes de numerales cardinales y ordinales.
- Conjunciones con su correspondiente categoría gramatical.
- Locuciones adverbiales y sustantivas (en este caso, se ha estimado oportuno introducir ejemplos reales de uso para contextualizar)
- Frases hechas
- *Phrasal verbs*
- Interjecciones
- Onomatopeyas y sonidos con los que se relacionan (*achoo – sneeze, slap – hit onface*)

El equipo de CÓDIGO ha codificado manualmente cada uno de los lemas en el *categorizador* desarrollado por el grupo de investigación «Text and Information Processing»(TIP, ULPGC) que ofrece una interfaz bastante intuitiva y fácil de usar. Esta aplicación permite categorizar cada palabra en función de su categoría y subcategoría gramaticales y permite incorporar los lemas de cada una de sus definiciones, así como sus sinónimos, antónimos, hiperónimos y antónimos. Además, permite asignar una notación diferente para cognados, falsos amigos y palabras deícticas.

Las pruebas de caracterización de informantes y de lectura, comprensión y traducción de fragmentos textuales ya están diseñadas y construidas, y en breve procederemos a realizarlas. El equipo de CÓDIGO prevé comenzar estas pruebas muy pronto y dará a conocer sus resultados en próximos eventos de divulgación científica.

6. Conclusiones

El proyecto CÓDIGO propone una línea de investigación novedosa en España. Entre sus principales aportaciones se encuentra el estudio de la *correfencialidad* entre términos y la aplicación de fórmulas de legibilidad a tres idiomas en el mercado profesional de la traducción. La herramienta en la que estamos trabajando es de utilidad inmediata y cuenta con el apoyo e interés de profesionales de reconocido prestigio de la Universidad de Massachusetts, la Comisión Europea, la Asociación de Empresas de Traducción (ACT) y agencias de traducción españolas como Hermes y Babel (Madrid) y Equus (Granada), entre otras. Desde una perspectiva científica, los procesos de lectura y comprensión se contemplan como fenómenos cognitivos complejos, por lo que CÓDIGO incluye una esmerada caracterización de los sujetos. Esto se complementa con un análisis de sus creencias implícitas, una vertiente innovadora que puede servir de punto de partida para investigaciones de mayor calado y que puede contribuir a establecer un paradigma experimental replicable fundamentado en la realidad psicológica. Finalmente —y aunque no está orientado a la docencia— la aplicación podrá utilizarse como herramienta para seleccionar textos de diversos grados de dificultad para asignaturas diversas en los programas universitarios de formación de traductores.

Referencias bibliográficas

Bolaños Medina, A. K. 2012. «Las pruebas psicométricas en la investigación empírica sobre los procesos cognitivos del traductor.» Cruces Colado, S., A. Luna Alonso, M. Del PozoTriviño y A. Álvarez Lugrís (eds.) *Traducir en la Frontera.* Granada: Atrio.

Bradac, J. J., R. A. Davies y J. A. Courtright. 1977. «The role of prior message context in evaluative judgments of high- and low-diversity messages» *Language and Speech* 20: 4, 295-307.

Caballero, A. 2002. «Entrevista con Juanjo Arevalillo». *La linterna del traductor* 1, 5-21.

Crossley, S. A., D. F. Dufty, P. M. McCarthy y D. S. MacNamara. 2007. «Toward a new readability: A mixed model approach.» MacNamara, D. S y G. Trafton (eds.) *Proceedings of the 29th annual conference of the Cognitive Science Society.* Austin (TX): Cognitive Science Society.

Douglas, J. Y. 2009. *A neuro-cognitive basis for readability.* Disponible en <http://writingcpr.com/page4/Neurocognition.html> (Última consulta, 21 de enero de 2012)

Dubay, W. H. 2004. *The Principles of Readability.* California: Impact Information.

Flesch, R. 1948. «A new readability yardstick» *Journal of Applied Psychology* 32, 221-233.

Folz, P. W., W. Kintsch y T. Landauer. 1998. «The measurement of textual coherence with Latent Semantic Analysis» *Discourse Processes*25: 2-3, 285-307.

— 1998. «The measurement of textual coherence with Latent Semantic Analysis» *Discourse Processes* 25: 2-3, 285-307.

Graesser, A. C., D. S. McNamara, M. M. Louwerse y Z. Cai. 2004. «Coh-Metrix: Analysis of text on cohesion and language» *Behavior Research Methods, Instruments, & Computers* 36, 193-202.

Kintsch, W. 1988. «The role of knowledge in discourse representation: A construction-integration model» *Psychological Review* 95, 163-182.

Landauer, T. K., P. W. Foltz y D. Laham. 1998. «An introduction to latent semantic analysis» *Discourse Processes* 25: 2-3, 259-284.

Lewis, J. R. 2006. «Effectiveness of various automated readability measures for the competitive evaluation of user documentation» *Proceedings of the human factors and ergonomics society 50th annual meeting*, 624-628.

McNamara, D. S., M. M. Louwerse y A. C. Graesser. 2002. *Coh-Metrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension*. Project funded by the Office of Educational Research and Improvement Reading Program. Disponible en <http://129.219.222.66:8080/SoletlabWeb/pdf/IESproposal.pdf> (Última consulta, 21 de enerode 2012)

Ministerio de Cultura. 2008. *La traducción editorial en España*. Centro de Documentación del Libro y la Lectura: Dirección General del Libro, Archivos y Bibliotecas.

Muñoz Martín, R. 2010. *Caracterización objetiva de la dificultad general de los originales (CÓDIGO)*. Memoria técnica del proyecto presentado a la convocatoria 2010 de ayudas para Proyectos de Investigación Fundamental no orientada del Ministerio de Ciencia e Innovación.

— 2009. «The waytheywere: Subject profiling in translation process research.» Mees, I., F. Alves y S. Göpferich (eds.) *Methodology, Technology and Innovation in Translation Process Research: A Tribute to Arnt Lykke Jakobsen*. Copenhague: SamfundslitteraturPress.

Muñoz Martín, R. y M. Castro Arce. 2012. «Caracterización objetiva de los textos originales.» Cruces Colado, S., A. Luna Alonso, M. Del Pozo Triviño y A. Álvarez Lugrís (eds.) *Traducir en la Frontera*. Granada: Atrio.

Neumann, S. y S. Hansen-Schirra. 2005. «The CroCo Project. Cross-linguistic corpora for the investigation of explicitation in translations» *Proceedings from the Corpus Linguistics Conference Series (PCLC),* 1:1.

Pegenaute Rodríguez, L. 2004. «La situación actual.» Lafarga, F. y L. Pegenaute (eds.) *Historia de la traducción en España*. Salamanca: Ambos Mundos.

Presas Corbella, M. y C. Martín de León. 2011. «Teorías implícitas de traductores principiantes. Una investigación cualitativa en traductología cognitiva» *Sendebar* 22, 87-112.

Royer, J. M., C. A. Green y C. M. Sinatra. 1987. «The sentence verification technique : A practical procedure for testing comprehension.» *Journal of reading* 30: 5, 414-422.