Revista Nebrija
de Lingüística Aplicada a la Enseñanza de las Lenguas

*Revista Nebrija de Lingüística Aplicada a la Enseñanza de Lenguas* (RNAEL) ISSN 1699-6569
Vol. 14 Núm. 29 (2020) doi: 10.26378/rnlael1429406
Recibido: 5/07/2020 / Aprobado: 12/10/2020


# Estudio comparativo de métodos de transcripción para corpus orales: el caso del español

## *Comparative study of transcription methods for spoken corpus: the case of Spanish*

**Marimar Rufino Morales**
*Université de Montréal, Canadá*
Mdm.rufino.morales@umontreal.ca

**RESUMEN**
Los avances tecnológicos han propulsado la metodología de investigación en transcripción. Los programas para corpus lingüísticos basados en modelos estadísticos y de aprendizaje profundo han mejorado las fases de alineación y anotación. En cambio, cuando se trata de transcribir el material, la carga interpretativa y la propia naturaleza de las conversaciones obstaculizan la automatización del proceso. De esta manera, la transcripción de entrevistas destinadas al estudio de la lengua oral se sigue haciendo con un reproductor y un teclado, y puede convertirse en uno de los aspectos más largos del procesamiento de datos. Sin embargo, en otros contextos profesionales, el reconocimiento automático del habla se emplea para transcribir de forma eficaz gracias a la colaboración humano-computadora. Las técnicas y estrategias difieren, pero todas tienen en común que estabilizan las fluctuaciones de las herramientas informáticas y son más rápidas que otros métodos. En este estudio se ha utilizado una de ellas, el rehablado *off-line* con las entrevistas del Corpus oral de la lengua española en Montreal. Se ha medido el tiempo empleado, así como la precisión y se ha comparado con el reconocimiento automático del habla y con la mecanografía. El rehablado *off-line* ha permitido el uso de un programa automático de dictado en su estado actual como herramienta para potenciar la transcripción de entrevistas en menos tiempo y con menos errores.

Palabras clave: rehablado, transcripción, reconocimiento automático del habla, programa de dictado, corpus oral

**ABSTRACT**
*Technological advances have propelled the research methodology in transcription. Language corpus tools based on statistical models and deep learning have improved the alignment and annotation phases. However, when it comes to transcribing the material, the conversation's interpretive load and nature themselves hinder automation of the process. That is why interviews used for studying spoken language are still transcribed with a player and keyboard, which can constitute one of the most time-consuming aspects of data processing. In other professional contexts, automatic speech recognition is used to transcribe effectively through human-computer collaboration. The techniques*

*and strategies may differ, but they all stabilize fluctuations in computing tools and are faster than other methods. In this study, the off-line respeaking method was used to transcribe the interviews of the Spoken Corpus of the Spanish Language in Montreal. Transcription times and accuracy were measured and compared with automatic speech recognition and typing. Off-line respeaking, using automatic speech-to-text software in its current state, proved to be the fastest and most error-free method for transcribing interviews.*

*Keywords: respeaking, transcription, automatic speech recognition, speech-to-text software, spoken corpus*

## 1. INTRODUCCIÓN

Los avances tecnológicos de las últimas décadas han propulsado la metodología de investigación en todos los ámbitos. También en lingüística, los corpus orales, herramienta esencial para el estudio del lenguaje humano, han aligerado las ineludibles fases previas de recopilación, transcripción y almacenaje de material. Alentados por el rendimiento que ofrecen numerosos ingenios de uso diario, no pocos investigadores han sometido las grabaciones de las entrevistas a uno o varios programas automáticos de dictado para llevar a cabo su transcripción. La gran oferta de programas para corpus lingüísticos donde se amalgaman transcripción, alineación, anotación (Kreuz y Riordan, 2018), identificación y análisis de patrones fomenta la expectativa de automatizar la transcripción ortográfica del material.

La Inteligencia Artificial y el aprendizaje profundo han mejorado, de manera insólita, la precisión de aquellos primeros programas de dictado que comenzaron a inundar nuestros mercados a principios de los 90. Su aplicación en el reconocimiento automático del habla (RAH) propulsó, a finales de 2016, un punto de ruptura: la paridad con el humano (Xiong et al., 2016). En este proceso de reconocimiento automático, basado en las redes neuronales artificiales, la voz se convierte en texto de forma casi instantánea y, sin embargo, intervienen innumerables operaciones de tratamiento del lenguaje natural. Es preciso analizar la señal, procesarla, codificarla, sintetizarla (Li Deng y Yang Liu, 2018). Sin el desarrollo de la microelectrónica (que aumentó la potencia y miniaturización de procesadores y memorias) y de la informática (que favoreció la expansión de las técnicas de comunicación y la interconectividad de las redes), nada hubiera sido posible (Mariani, 2002). Casi un siglo de numerosos experimentos e ingentes esfuerzos de colaboración internacional separan a los asistentes virtuales como Siri o Cortana del vocoder, aquel sintetizador de voz creado en los laboratorios de Bell (Dudley, 1958) y presentado en la Exposición Universal de Nueva York en 1939.

Unos datos tan prometedores merecen ser aclamados, pero también han de ponerse en perspectiva; de hecho, las transcripciones destinadas al estudio de la lengua hablada las siguen haciendo los humanos (Durand, 2017). A las numerosas decisiones que hay que tomar durante el proceso de transcripción, y que lo alejan de un acto puramente mecánico (Ochs, 1979), se suman las limitaciones propias de esta tecnología de RAH. Por una parte, el entrenar modelos acústicos con aprendizaje profundo nos proporciona datos falsamente reales (Saon et al., 2017). Por ejemplo, si en el laboratorio 'enseñamos' a un programa de

RAH abasteciéndolo con cuarenta entrevistas semidirigidas extraídas de un corpus que contiene sesenta, realizadas todas ellas por la misma persona, quien a su vez ha formulado siempre más o menos las mismas preguntas, probablemente que la transcripción de las veinte entrevistas restantes con ese mismo programa una vez entrenado, sea excelente. Por otra, a pesar de que muchos científicos trabajan intensamente para mejorar las tecnologías del habla (Zweigenbaum et al., 2020), ante la conversación espontánea o la variación, los programas de RAH aún no son estables (Ravanelli et al., 2018). En definitiva, la transcripción automática de la lengua hablada sigue enfrentándose a las limitaciones de la tecnología. Programas tales como *Transana*, *Soundscriber*, *Transcriber* o *Audacity* funcionan bien como herramientas para etiquetar la transcripción, pero para transcribir, todavía no han suplantado al humano (Revuelta Domínguez y Sánchez Gómez 2005). Los de mayor utilidad son aquellos que, basados en modelos estadísticos y de aprendizaje profundo, ayudan en las fases de alineación y anotación de corpus (Yadav et al. 2018). También en español estos programas se ocupan menos del primer nivel de representación, la transcripción ortográfica, si bien automatizan con éxito la codificación, por ejemplo, de características suprasegmentales de la prosodia, tales como la entonación para la transcripción fonética (Elvira-García et al., 2015).

Y, sin embargo, tenemos constancia del uso eficiente y eficaz de la tecnología de RAH y de los programas de dictado automático, para transcribir, en varios contextos profesionales donde, desde hace dos décadas, se han venido desarrollando métodos basados en la colaboración humano-computadora. Las técnicas y estrategias de dictado difieren según el contexto, pero todas tienen en común que estabilizan las fluctuaciones de las herramientas informáticas empleadas. Nos hemos centrado en el rehablado, que se usa en televisión (Romero-Fresco, 2011), por el paralelismo entre la oralidad que está presente en la mayoría de los escenarios o situaciones de los medios audiovisuales y la que caracteriza los datos que se recopilan para crear un corpus de la lengua hablada. Hemos empleado los programas de dictado con una adaptación del rehablado, logrando, al igual que en televisión, optimizar el proceso de transcripción (Rufino Morales, 2020).Para abordar la cuestión, hemos confrontado el rehablado con los otros métodos de transcripción de corpus orales para la investigación lingüística. Presentamos aquí los resultados de los datos recopilados durante el primer taller de rehablado *off-line* que se organizó el pasado mes de enero en la Sección de estudios hispánicos de la Universidad de Montreal. Se ha comparado el rehablado con el reconocimiento automático del habla (RAH) y con la mecanografía, en cuanto al tiempo empleado, así como la precisión resultante al transcribir entrevistas del Corpus oral de la lengua española en Montreal (COLEM), (Pato dir.). Es un corpus fruto de una serie de entrevistas semidirigidas grabadas en entorno natural, de una hora de duración aproximadamente y estructuradas en función de un protocolo de encuesta común. El estilo que recoge el COLEM es el conversacional y familiar (habla espontánea). Refleja la situación del español en la Región Metropolitana Montreal, única en el mundo, por el contacto mantenido con el francés y el inglés, pero también por el contacto de todas las variedades del español, sin que una se imponga sobre las demás. En este contexto se encuentra, además, la mayor concentración de latinoamericanos de todo Canadá (Pato, 2017).

Este trabajo forma parte de una investigación cuyo objetivo es describir el perfil *ad hoc* del rehablador *off-line* para optimizar la transcripción de grabaciones, al menos hasta que los programas informáticos de reconocimiento automático del habla permitan cederles el testigo. Primero revisaremos en qué ámbitos se transcribe de forma eficaz con ayuda de programas de dictado; comprobaremos que la investigación cualitativa también ha intentado usarlos sin llegar a adoptarlos; finalmente, ofreceremos una prueba de que es

posible utilizarlos para potenciar la transcripción de corpus orales de la lengua hablada siempre y cuando se empleen las estrategias adecuadas. De seguro, las aproximaciones de este manuscrito retendrán la atención de otros investigadores que trabajan con textos orales.

## 2. TRANSCRIPCIÓN CON UN PROGRAMA DE DICTADO

Convertir la voz a texto es una necesidad presente en todas las épocas y en distintas esferas de la actividad humana, ya sea de forma intralingüística o interlingüística. No obstante, desde el punto de vista diatópico, las herramientas (o la forma de usarlas) difieren. Si optamos por un método al alcance de todos, un reproductor y un teclado, nos enfrentaremos al desfase entre la velocidad del habla y la de la escritura. "La velocidad de habla o tempo de elocución es la rapidez con que una persona articula las palabras a lo largo de su discurso. Para determinar la velocidad, se computa la cantidad de palabras que emite en un periodo de tiempo. El resultado se expresa, generalmente, en palabras por minuto; en español se calcula que una velocidad normal oscila entre las ciento cincuenta y las doscientas palabras por minuto" (AA.VV. 2008).

La velocidad óptima de producción oral se sitúa entre ciento setenta y ciento noventa palabras por minuto (Rodero Antón, 2016), pero el habla espontánea no se ciñe a este canon, pudiendo ser más elevada. En cuanto a la velocidad de escritura, alguien sin formación teclea entre quince y veinticinco palabras por minuto (Ainsworth, 1988); la media de una persona experimentada se sitúa entre ochenta y noventa palabras por minuto (Moro Vallina, 2010: 7). Además de esta diferencia entre la velocidad a la que hablamos y la velocidad a la que escribimos, al transcribir una grabación habrá que detenerla, retroceder, volver a ponerla en marcha, escuchar de nuevo, etc.

Podemos sustituir el teclado por el RAH, lo que además restará subjetividad a la transcripción (Tatham y Morton, 2005: 372). Pero entonces, tendremos que enfrentarnos a los errores provocados por las variantes geográficas (Winata et al., 2020) o sociodemográficas; también habrá que borrar y reescribir una representación más cercana a la actuación del hablante cada vez que el programa de dictado le haya atribuido una correspondencia que figura en su léxico a palabras fragmentadas, mal pronunciadas, solapadas o inexistentes. De manera que la automatización del proceso termina siendo más tediosa que la clásica transcripción mecanografiada, sobre todo si sumamos el tiempo que puede tomar aprender las nociones básicas de la herramienta que haya decidido emplearse. La comunidad científica es unánime: fuera del laboratorio, la tecnología de RAH aún no está a punto para convertir con exactitud la voz en texto en situaciones reales, donde, por ende, los ruidos y la reverberación afectan a la calidad de la señal de voz emitida por el hablante (Lu et al., 2020).

Una forma de estabilizar los resultados del RAH consiste en utilizar estrategias de repetición que, emulando el modo de corrección privilegiado en el diálogo humano-humano (Brinton et al., 1986), pero aplicado a la comunicación humano-computadora, consiguen optimizar la transcripción en tiempo real con ayuda de un programa de dictado. Es lo que acometen los siguientes métodos.

### 2.1 El rehablado

El rehablado surgió para paliar las limitaciones de la tecnología de RAH a la hora de subtitular en tiempo real en televisión (Utray Delgado et al., 2015; Lambourne, 2007). El rehablador, instalado en un entorno exento de ruido, a la vez que escucha la emisión a subtitular, la repite o parafrasea verbalmente a través de un micrófono conectado a un programa de dictado, de manera que genere una transcripción legible de forma automática. El texto resultante es enviado al codificador de subtítulos (Brousseau et al., 2003) y la emisión se hace accesible, para aquellas personas que no pueden oírla, casi simultáneamente.

La subtitulación en directo consiste en emitir la transcripción del contenido sonoro de forma simultánea a la difusión de un programa audiovisual. Los subtítulos enviados en directo pueden haber sido preparados con antelación o en tiempo real (ACR/CAB 2012). Deben incluir además de las palabras, la prosodia (entonación, acento, ritmos), los efectos sonoros, las señales musicales y cualquier otra información de la banda sonora pertinente, de modo que la lectura de los subtítulos y la escucha del audio proporcionen una comprensión análoga de dicho programa (Ivarsson, 1992).

Los subtítulos y la audiodescripción son instrumentos vitales para garantizar el derecho de acceso a la comunicación audiovisual de todas las personas por igual (ONU, 1994); la aplicación de dicho principio de igualdad queda asegurada a través de medidas legales a nivel de cada país. Así, por ejemplo, desde 1995, el Gobierno de Canadá comenzó instando a los teledifusores tanto de lengua francesa como de lengua inglesa a aumentar el contenido subtitulado, hasta que en 2007, los obligó a subtitular el 100% de la programación (CRTC, 2007).

Los primeros subtítulos en directo para televisión se enviaron por teletexto; para su confección, se probaron distintas maneras y también varios teclados de estenotipia (Hawkins y Robinson, 1979; Tanton, 1979). Por aquel entonces, se trataba del método más extendido para transcribir en tiempo real. Pero la estenotipia necesita una instrucción intensiva de más de dos años. Por otra parte, existen distintos sistemas, máquinas y programas informáticos con sendos métodos, difícilmente intercambiables (Manrique Fuero, 2016). Por ejemplo, la compañía americana Stenograph (www.stenograph.com) comercializa varias máquinas de estenotipia: Diamante, Élan Mira, Stentura. Familiarizarse con un teclado toma mucho tiempo. En Quebec, la formación profesional de transcriptores se da en l'École de sténographie judiciaire de Québec. Los estudiantes deben procurarse una de las dos máquinas sugeridas con la que podrán aprender el programa *Case CATalyst*. Pero al cabo de doscientas setenta horas de clase y noventa más de prácticas laborales, los diplomados solo habrán utilizado un teclado y un programa, y difícilmente podrán trabajar con otro.

De manera que, para subtitular en directo en televisión, desde el primer momento se buscaron alternativas más rentables que la estenotipia. Fue así como surgió la idea de usar los programas de RAH potenciados por la interacción humano-computadora (Damper et al., 1985; McCoy y Shumway, 1979). El éxito del rehablado on-line a la hora de subtitular en directo y en tiempo real en televisión reside en que este método requiere menos personal (un rehablador subtitula un programa de treinta minutos; dos rehabladores por turnos hacen accesibles ocho horas continuas de programación). Además, es una técnica que se domina en poco tiempo, salvando la brecha entre la formación en empresa (de dos a cuatro

semanas) y los módulos y cursos académicos disponibles (hasta seis meses) (Bernabé Caro et al., 2019).

Las ventajas que ofrece el rehablado *on-line* a la hora de producir subtítulos en directo en televisión han extendido su uso a muchos otros ámbitos. En la actualidad, tanto en universidades, colegios, actos públicos, eventos de masas (conferencias, fórums, seminarios), reuniones, teleconferencias, podcasts, programas de radio, teatros, museos, iglesias (Moores, 2016; Romero-Fresco, 2012), el rehablado on-line proporciona accesibilidad en tiempo real del contenido sonoro emitido en directo, de forma presencial o por internet (portátil, tableta, teléfono).

Asimismo, muchas de las subtituladoras que cuentan con la infraestructura de rehablado *on-line* aprovechan sus recursos para acelerar la transcripción de programas 'enlatados' o en diferido (FCC, 2014). Para transcribir en tiempo real documentos audiovisuales grabados con antelación se emplean estrategias particulares. El objetivo no es mejorar la inteligibilidad del enunciado obtenido por el programa de dictado (que no va a ser leído por los televidentes de forma simultánea), sino reducir el trabajo de edición posterior de los subtítulos.

Las estrategias empleadas en el rehablado *off-line* sirven para crear un documento lo más cercano posible a su versión definitiva. Este es el método que estoy usando en mi investigación, donde he adoptado una nomenclatura que distingue dos formas de rehablado, según las circunstancias en las que se produce la transcripción: *on-line* y *off-line* (Lindsay y O'Connell, 1995; Ferber, 1991).

### 2.2. La escritura de voz

En otros contextos donde también se requiere la transcripción de un discurso hablado, como en el campo jurídico y en la administración pública, el primer método eficaz que se empleó para transcribir al mismo tiempo que se produce el discurso oral fue la taquigrafía. De él derivan la estenotipia y la escritura de voz. Mientras que la taquigrafía (también llamada estenografía) emplea trazos breves, abreviaturas y caracteres especiales para representar las letras, palabras y frases, la estenotipia utiliza un teclado reducido basado en las sílabas (Núñez Hidalgo y Ramos Villajos, 2010).

La escritura de voz (*voice writing)* es otra forma de transcribir; pero, en este caso, como su nombre indica, mediante la voz. Para que al dictar, el escritor de voz (*voice writer*) no moleste a las otras personas presentes en la reunión –y para atenuar el ruido ambiente–, se emplea un silenciador de voz. La idea le vino a Horace Webb, un transcriptor en el tribunal de Chicago que, en los años 40, insertó un micrófono en una caja de cigarrillos, y después en una lata de café. Amortiguaba el sonido mientras repetía y grababa lo que oía en el tribunal y conseguía transcribir palabra por palabra el discurso original. Los escritores de voz adoptaron los programas automáticos de dictado y un nuevo método de transcripción, que se extendió rápidamente en los años 2000. El oficio de transcriptor oficial (*verbatim reporter*) en Estados Unidos tiene dos especializaciones: la estenotipia y la escritura de voz. La escritura de voz cuenta con dos ramas con formación y acreditación propia: para transcribir en juzgados (*court and real time reporting*) o en televisión (*captioning*) (NVRA, 2008).

Así como el rehablador, el escritor de voz también transcribe a partir de grabaciones, ya sea la defensa de un abogado, el veredicto de un juez o un testimonio. Por la naturaleza oficial de la tarea, el resultado final debe corresponderse palabra por palabra con el discurso original; ahora bien, ese mismo carácter resta espontaneidad a la interacción discursiva.

### 2.3. El dictado médico

Numerosas actividades en el ámbito de la salud requieren transcripción (Pollard et al. 2013). Hasta el giro tecnológico de los 90, los médicos y especialistas que necesitaban transcribir se lo encargaban a taquígrafos o administrativos. Desde entonces, cada vez más profesionales han ido integrando los programas de dictado, ya sea para elaborar la historia digital del paciente (Johnson et al., 2014), redactar un informe en radiología, endocrinología, psiquiatría o en patología quirúrgica, entre otros.

Por lo general, los médicos que recurren al dictado envían el texto generado a un administrativo para su revisión, antes de darle el visto bueno final. Pero en otros casos, el propio médico edita su transcripción generada con el programa de RAH. Sea cual sea el método empleado, la revisión es sumamente importante (Zhou et al., 2018) porque, a pesar de que los programas de dictado constituyen un buen punto de partida para transcribir (Edwards et al., 2017), en el ámbito médico también se constatan las mismas deficiencias del RAH que en otros contextos (Blackley et al., 2019). Por ende, la automatización del proceso plantea un problema ético: altas tasas de precisión no son sinónimo de resultados clínicamente seguros; en ocasiones han llegado a poner en peligro la seguridad del paciente (Hodgson y Coiera, 2016).

Estudios recientes centrados en la tipología de los errores durante el uso de aplicaciones de reconocimiento automático del habla en el campo médico confirman que: i) los mayores errores se producen en las conversaciones casuales (Chiu et al., 2018); ii) los resultados mejoran considerablemente con programas de vocabulario especializado y si se aprovechan bien los atajos (*shortcuts*) que ofrece el programa (Edwards et al., 2017) y iii) la precisión aumenta con una metodología estandarizada (Blackley et al., 2019).

Y, sin embargo, el uso de estrategias para mejorar los resultados de los programas de dictado no parece haber pasado de iniciativas aisladas. MacLean, Meyer *et al.* (2004: 115) emplearon "*an oral transcriptionist to act as an intermediary*" en la transcripción de entrevistas con *Dragon NaturallySpeaking*. El método *leasen and repeat* se usó en un experimento en la universidad de Auckland para entrenar a sus participantes a servirse de *Dragon* y a repetir verbalmente las grabaciones (Park y Zeanah 2005). Sirvió de inspiración, primero, para la técnica de transcripción vocal o *Voice Transcription Technique* (VTT) (Matheson, 2007) y, posteriormente, para otra llamada *Embodied Transcription* (ET), porque se encarna a los entrevistados al repetir sus palabras para acelerar el proceso de transcripción con un programa de dictado (Brooks, 2010).

## 3. LA TRANSCRIPCIÓN EN LA INVESTIGACIÓN LINGÜÍSTICA Y CUALITATIVA

La revisión de las actas de eventos internacionales como los organizados por la European Language Ressources Association (www.elra.info) o la International Speech Comunication Association (www.isca-speech.org), nos confirma que la transcripción de la lengua hablada, larga y costosa, es más difícil cuanto más detallado sea el nivel de transcripción (Adolphs y Knight, 2010) y que los grandes corpus se siguen transcribiendo de forma manual (Niemants, 2018; Gadet et al., 2012) o se completan de forma manual (www.isip.piconepress.com). Transcribir una hora de entrevista puede tomar entre cuatro y sesenta horas (Markle et al., 2011). Como en los demás contextos donde se requiere una transcripción *off-line*, también aquí se han puesto a prueba los programas de dictado para acelerar la tarea de transcripción de las grabaciones, laboriosa y lenta *per se*. Necesitan ser

más rápidos que las otras formas de transcribir al alcance; pero también, en su adopción como herramienta, juegan un papel decisivo la precisión de los resultados obtenidos.

Aunque la investigación cualitativa se ha ocupado más de la metodología para producir una transcripción que de la mecánica empleada durante el proceso, varios investigadores han estudiado el uso de programas de dictado para transcribir datos grabados de forma automática. Los resultados obtenidos contenían tantos errores que nadie ha llegado a considerar el RAH como una alternativa seria a la transcripción manual. Argumentan que, dependiendo de la velocidad a la que la persona escriba, el tiempo de aprendizaje y entrenamiento del programa no compensa, a menos que se explore un método de optimización de las herramientas informáticas (Evers, 2011). En este sentido, dos trabajos han retenido nuestra atención.

Tilley investigó el principio de co-construcción de la transcripción a través de la experiencia de personas contratadas en universidades canadienses para transcribir grabaciones académicas (2003). Reportó que uno de los transcriptores utilizaba un programa de RAH y documentó cómo este se fue habilitando, sobre la marcha, con estrategias para mejorar la precisión. Según Tilley, el transcriptor reproducía, sin saberlo, el efecto fantasma (*ghosting*) que habían acuñado Frogg y Wightman (2000).

Por su parte, Johnson dictó una entrevista mientras la iba escuchando, luego de haber entrenado el programa *MacSpeech Dictate* (2011). El dictado de la entrevista (17:38 minutos) le tomó treinta minutos cincuenta y seis segundos, ya que había disminuido la velocidad para lograr repetir todo el contenido de la grabación. La transcripción obtenida tenía 96,4% de precisión y necesitó veintinueve minutos para corregirla. Después llevó a cabo la transcripción de la misma entrevista de forma manual, es decir, escuchando y tecleando al ordenador. Le tomó treinta y nueve minutos y siete segundos, y obtuvo 98% de precisión. Necesitó menos de doce minutos para corregirla. Concluyó que los programas de dictado no son más rápidos ni más precisos que el método manual. La validez de la prueba puede cuestionarse (entre otras cosas porque cuando tecleó la entrevista ya la había oído al menos tres veces), pero también apoya nuestro postulado: no habrá optimización del RAH si no se emplean las estrategias adecuadas durante la repetición.

Lo mismo ocurre con otros experimentos que se han llevado a cabo para medir el potencial de los programas de dictado utilizados de forma automática o potenciados con distintas técnicas de repetición a la hora de transcribir grabaciones *off-line*. Comparan transcripciones obtenidas, entre otras, de forma manual (mecanografía y/o estenotipia) con transcripciones obtenidas con la técnica de *leasen and repeat* (Johnson, 2011; Matheson, 2007; Park y Zeanah, 2005) o con rehablado *on-line* (Matamala et al., 2017; D'Arcangelo y Cellini, 2013). Ninguna de las técnicas de dictado empleadas ofrece una neta ventaja con respecto a los demás métodos de transcripción.

Teniendo en cuenta que la transcripción de entrevistas para la investigación, basada en la literalidad (Davidson 2009), debe cumplir con unos requisitos, reflejados en un protocolo, dependiendo del uso que se le vaya a dar, hemos optado por aplicar nuestras propias estrategias. Fuera de los trabajos que estamos realizando, no tenemos constancia de ningún otro que mida estrategias específicas de rehablado para optimizar la transcripción *off-line* del habla espontánea con fines lingüísticos.

## 4. METODOLOGÍA

Los datos que analizo provienen del primer taller de rehablado *off-line* para transcribir corpus orales de la lengua hablada organizado en la Universidad de Montreal. Se trata de una actividad vinculada al proyecto El español en Montreal y el COLEM (CERAS-2014-15-159D) que, por estar dirigida a estudiantes de segundo y tercer ciclo de la Sección de estudios hispánicos, nos garantizaba el nivel de lengua necesario a la hora de transcribir y revisar entrevistas con las variaciones diatópicas expresadas por los hispanohablantes de la Región metropolitana de Montreal.

Los diez participantes (siete hombres y tres mujeres) estaban familiarizados con los corpus orales de la lengua, pero ninguno había rehablado antes. En este trabajo, se emplean los datos de los seis que completaron todas las actividades.

Se organizaron dos jornadas completas. De la primera se han extraído los datos sobre la mecanografía. Se ha medido tanto la cantidad como la calidad de la transcripción obtenida por cada participante durante treinta minutos: primero, al revisar una entrevista rehablada por un profesional; segundo, al mecanografiar la continuación de la entrevista. Somos conscientes de la ventaja que puede suponer, en el momento de mecanografiar, el haber tenido ya contacto con parte de la entrevista rehablada, pero en nuestra elección primó mantener el nivel de lengua. Se utilizó una entrevista del COLEM realizada en 2018 con un informante procedente de Moca, República Dominicana de treinta y siete años y treinta y tres en Montreal (E32) (60:54 minutos). Al inicio del taller, se dedicó un tiempo para que los participantes se familiarizaran con las pautas de transcripción del COLEM y con los atajos de teclado del reproductor multimedia. Seguidamente se les suministró un texto (E32b) correspondiente a la primera parte de la entrevista E32 (29:50 minutos) rehablada por un profesional. Disponían de treinta minutos para revisarlo respetando el protocolo del COLEM. También se midió la velocidad a la que los participantes son capaces de teclear (10fastfingers.com).

Con objeto de obtener datos precisos sobre el rehablado, en el segundo día, proporcionamos la grabación de una entrevista distinta del COLEM a cada participante. Debían rehablarla y editar la transcripción resultante, luego de una iniciación a las principales estrategias del rehablado *off-line*. Aparecen aquí los datos de los dos primeros participantes que completaron la tarea. L1 rehabló parte de una entrevista realizada en 2019 con una informante de Santiago de Chile de cuarenta y siete años, y nueve en Montreal (E5) (30:27 minutos). L2 rehabló la entrevista completa realizada en 2019 con un informante de La Estrella, Chiriquí, Panamá, de cincuenta y ocho años, y treinta en Montreal (E27) (72:32 minutos).

Para determinar si las variantes del español y la experiencia del rehablador pueden repercutir en los resultados, hemos cotejado los datos obtenidos con L1 y L2 con dos entrevistas rehabladas por un profesional (L0). La entrevista E4 se realizó en 2019 con un informante procedente de Santiago de Chile de cincuenta años y once en Montreal (106:41 minutos); E26 se hizo en 2019 con un informante de la Ciudad de Panamá de cincuenta y un años, y treinta en Montreal (69:20 minutos).

En el marco de mi investigación, las entrevistas del COLEM se han transcrito con distintos programas de dictado, y se han comparado las transcripciones automáticas con las rehabladas. Para el presente artículo, las transcripciones se han realizado con el programa *Dragon Naturally Speaking 13*, por ser representativo de la media de los programas comparados. Presento los resultados obtenidos de la transcripción automática de las entrevistas para cotejarlos con los otros dos métodos.

Todas las transcripciones han sido tratadas en el Centre de Recherche Informatique de Montréal con la herramienta *Align-text* de Kaldi (kaldi-asr.org). Se ha medido el índice de precisión, expresado en porcentaje, según la fórmula del National Institute of Standards and Technology que alinea cada texto con su correspondiente versión final (T0). El porcentaje de precisión se ha calculado dividiendo la suma de borrados (B) (cuando falta algo que figura en el T0), sustituciones (S) (por ejemplo, donde en lugar de adonde) e inserciones (I) (cuando se añade algo que no figura en el T0), por el número de unidades de referencia (N). A su vez, (N) corresponde a la suma de (B + S + C), siendo (C) las unidades correctas.

La versión final (T0) hace referencia a una transcripción corregida según el protocolo del COLEM; esto es, el T0 corresponde a un texto que, sea cual sea el método empleado para obtenerlo, antes de ser sometido a una lectura final por el director del proyecto, debe ser comparado con su audio por un revisor. En este sentido, la revisión de transcripciones de entrevistas puede compararse a la revisión en traducción, ya que ambas requieren revisión y relectura. La revisión comparada del texto meta con el de origen (en nuestro caso, la grabación de la entrevista) es necesaria, entre otras cosas, para comprobar que no ha habido omisiones durante la transferencia (AENOR, 2006).

Por otra parte, no existe un conjunto de reglas y criterios universales para establecer la calidad de una transcripción, dependerá del propósito al que vaya destinada y siempre será una herramienta interpretativa (Lapadat, 2000). Para esta investigación, el índice de precisión no solo tiene en cuenta caracteres o palabras. Le hemos otorgado el mismo valor a palabras, signos de puntuación y otras convecciones del COLEM que forman parte del protocolo: omisión de todos los nombres propios representados con [NP]; identificación del informante con [I:] y del entrevistador con [E:]; inclusión de rasgos suprasegmentales: entonación, ritmo, tono, pausas; elementos paralingüísticos como aplauso, beso y demás recursos cinéticos audibles; reproducción exacta de las alteraciones morfosintácticas. Se trata de una forma de cuantificar el hecho de que sea cual sea la acción a realizar para corregir, siempre habrá que detener el audio y luego editar el texto; en definitiva, siempre supone tiempo.

## 5. RESULTADOS Y ANÁLISIS

### 5.1. La mecanografía frente al rehablado

Los datos de la Figura 1 comparan la mecanografía frente al rehablado. Consta de tres bloques. Primero se midió la velocidad a la que cada participante es capaz de teclear durante un minuto, expresada en número de palabras por minuto (ppm), y el porcentaje de precisión. Después, a partir de la entrevista E32b, medimos los minutos y el número de palabras mecanografiados (T5) durante treinta minutos, así como el número de unidades contabilizadas y la precisión. En el tercer bloque figuran los minutos, las palabras, el número de unidades contabilizadas y la precisión de la transcripción T5, que revisó cada participante, durante treinta minutos; corresponde a la entrevista E32a rehablada por un profesional. En la última línea se han calculado los promedios correspondientes a un minuto.

| Participante | L1 | L2 | L3 | L4 | L5 | L6 | Media 1 minuto |
|---|---|---|---|---|---|---|---|
| **Prueba de mecanografía: ppm** | 79 | 64 | 43 | 71 | 71 | 52 | 63,33 |
| **Prueba de mecanografía: precisión (%)** | 99,50 | 92,80 | 99,09 | 97,51 | 95,16 | 89,90 | 95,66 |
| **T5-E32b: minutos de grabación** | 02:07 | 02:02 | 01:38 | 02:30 | 02:32 | 04:09 | 02:29 |
| **T5-E32b: palabras** | 342 | 362 | 250 | 445 | 272 | 608 | 12,66 |
| **T5-E32b: unidades** | 557 | 508 | 393 | 635 | 392 | 1003 | 19,38 |
| **T5-E32b: precisión (%)** | 58,28 | 76,06 | 66,49 | 63,80 | 68,38 | 60,16 | 65,53 |
| **T6-E32a: minutos de grabación** | 13:46 | 06:23 | 07:47 | 04:52 | 23:35 | 15:46 | 12:01 |
| **T6-E32a: palabras** | 1809 | 1187 | 1387 | 980 | 4355 | 2643 | 68,67 |
| **T6-E32a: unidades** | 2735 | 1756 | 2059 | 1434 | 6394 | 3939 | 101,76 |
| **T6-E32a: precisión (%)** | 78,54 | 80,15 | 78,19 | 73,40 | 81,73 | 79,61 | 78,60 |

Figura 1. Prueba de mecanografía / transcripción mecanografiada (T5-E32b) / transcripción rehablada (T6-E32a)

Los participantes teclean una media de 63,33 ppm, cercana a la de un mecanógrafo experimentado (Moro Vallina, 2010). Sin embargo, al mecanografiar la entrevista E32, obtienen una media de 12,66 ppm. Esta diferencia puede explicarse, en primer lugar, porque en las pruebas de mecanografía como 10fastfingers.com, las palabras se suceden separadas por espacios, sin ningún otro signo de puntuación ni mayúsculas; además, a la hora de mecanografiar la entrevista hay que considerar el tiempo empleado en manipular el audio y el tempo del habla. Con la entrevista rehablada los participantes no necesitan teclear cada palabra; el número de palabras revisadas asciende a una media de 68,67 ppm.

Mecanografiando la entrevista E32b durante treinta minutos, los participantes solo avanzan una media de 2:29 minutos; mientras que, en el mismo tiempo, con la entrevista E32a rehablada —que además contiene ya una revisión—, la media de transcripción obtenida asciende a 12:01 minutos. Esto es, aun añadiendo el tiempo de rehablado, que equivale a 1X1 (un minuto de grabación se rehabla en un minuto), esta forma de transcribir sigue siendo más rápida. Podríamos deducir que para obtener la transcripción completa de la entrevista E32 (60:54 minutos) mecanografiándola se necesitarían unas trece horas, mientras que a partir del rehablado, la media sería de menos de tres horas. Aun sumando el tiempo empleado al rehablarla, la transcripción T6 tomó menos tiempo que la mecanografiada; pero, además, es más precisa.

Porque si la rapidez a la que obtengamos la transcripción de la entrevista cuenta, la calidad de su contenido es primordial. En el ámbito del COLEM, para que una transcripción sea de calidad, además de ser fiel al contenido, tendrá que responder rigurosamente al estándar de su protocolo de transcripción. Por eso, aunque durante la prueba de mecanografía se midió la precisión, la media, que solo cuenta palabras, no nos sirve. Nos fijaremos en los porcentajes de precisión de las columnas 7 y 11, resultantes de la comparación de las transcripciones mecanografiadas (T5) y rehabladas (T6) con sus correspondientes T0. Estos porcentajes sí representan un baremo de calidad porque al

comparar unidades en lugar palabras tendrán en cuenta todos los elementos que debe reflejar la transcripción del COLEM.

La precisión media al mecanografiar fue de 65,53%, inferior a la de la transcripción rehablada: 78,60%. He aquí algunos ejemplos de los errores y cómo se han contabilizado: (I) por inserción, (S) por sustitución.

| T0 | T5 / T6 |
|---|---|
| Para identificar entrevistador se emplea la etiqueta [E:]<br>seguida de tabulador (sin espacio) | • [E] (S)<br>• seguida de espacio (I),<br>• seguido de dos puntos (I)<br>• y uno o varios espacios (I) |
| Para identificar al informante se emplea la etiqueta [I:]<br>seguida de tabulador (sin espacio) | • [I] (S)<br>• seguida de espacio (I)<br>• seguido de dos puntos (I)<br>• y uno o varios espacios (I) |
| Todos los nombres propios se sustituyen por [NP] | • Aparece el nombre (S) |
| [RISAS]<br><br>[SOLAPAMIENTO]<br><br>[EXPIRACIÓN]<br><br>[RUIDO] | • (RISAS) (S)<br>• [RISA] (S)<br>• [Soplamiento] (S)<br>• (SOLOPAMIENTO) (S)<br>• [RESOPLIDO] (S)<br>• [RESOPLIDO] (S)<br>• [toca] (S) |
| La aspiración de vocales y consonantes no se representa:<br>dieciséis<br>trabajadores | • dieciséi (S)<br>• trabajadore (S) |
| Representación completa de palabras cortadas:<br>conserv-<br>llega- | • conservaron (S)<br>• llegaste (S) |

Figura 2. Errores de formato en las transcripciones

Los errores se han detectado en ambas versiones, pero en mayor proporción en la mecanografiada; por tanto, la automatización del proceso reduce la parcialidad y la fuerte carga interpretativa inherentes a la transcripción.

Por último, la observación combinada de tiempo y precisión no permite pensar que exista correlación entre la velocidad y la calidad. L5, fue quien más avanzó (23:35 minutos); pero también fue quien obtuvo la mayor precisión (81,73%). Con el siguiente mejor porcentaje (80,15%), L2 solo avanzó 6:23 minutos.

### 5.2. El RAH frente al rehablado (experto / principiante)

En la figura 3 se comparan transcripciones obtenidas con el RAH (T1) con transcripciones rehabladas (T2). Se han medido los índices de precisión de cuatro entrevistas realizadas a dos chilenos (E4 y E5) y dos panameños (E26 y E27). El rehablado de E4 (Chile) y E26 (Panamá) los hizo un experto (L0). L1 rehabló la entrevista E5 (Chile) y L2 rehabló la E27 (Panamá); para ambos fue su primer rehablado. Las ocho transcripciones se han obtenido con el programa de dictado *Dragon NaturallySpeaking 13*.

| Identificación | Transcripción | Precisión | Duración |
|---|---|---|---|
| T1-E4 | RAH | 12,24% | 106:41:00 |
| T2-E4 | L0 (experto) | 87,75% | 106:41:00 |
| T1-E5 | RAH | 36,93% | 90:59:00 |
| T2-E5 | L1 | 62,10% | 30:27:00 |
| T1-E26 | RAH | 25,54% | 69:20:00 |
| T2-E26 | L0 (experto) | 74,46% | 69:20:00 |
| T1-E27 | RAH | 5,53% | 72:32:00 |
| T2-E27 | L2 | 67,77% | 72:32:00 |

Figura 3. Transcripción automática / Transcripción rehablada

Ninguna de las transcripciones automáticas (T1) consigue la paridad con el humano (Xiong et al. 2016). La precisión de todas las transcripciones automáticas está por debajo de las rehabladas (T2), confirmándose dos premisas iniciales: por una parte, para transcribir la lengua hablada, el RAH aún no permite la completa automatización del proceso con un programa de dictado automático (Ravanelli et al. 2018); por otra, el rehablado mejora los resultados del RAH de los programas de los programas de dictado (Utray Delgado et al. 2015; Romero-Fresco 2011; Lambourne 2007).

Los errores léxicos reflejan cómo afecta la variación del español en los resultados: inserciones (I), borrados (B), sustituciones (S) y equivalencias correctas (C). En los siguientes ejemplos aparece la palabra tal y como se pronuncia en el texto de origen que es la grabación (T0) con su correspondiente representación en la transcripción T1 (RAH) y T2 (rehablado).

| T0 | T1 | T2 |
|---|---|---|
| amerindiez | abrir indios (S) + (I) | (B) |
| citación | situación (S) | citación (C) |
| outsourcing | (B) | outsourcing (C) |
| intalando | instalando (S) | instalando (S) |
| candenciosa | (B) | ambiciosa (S) |
| quartier | catión (S) | cartilla (S) |
| marché | marche (S) | marche (S) |
| incribían | vivía (S) | inscribían (S) |
| úneco | único (S) | único (S) |
| tranporte | transporte (S) | transporte (S) |
| afordable | (B) | razonable (S) |
| contrucción | confusión (S) | construcción (S) |
| indostanes | en dos canes (I) + (S) + (I) | (B) |
| bombardios | un (S) | bombardeos (S) |
| distes | tristes (S) | diste (S) |
| cargastes | (B) | cargas tests (S) + (I) |
| meritaba | meditaba (S) | meditaba (S) |
| noventicinco | (B) | 95 (S) |
| quebecuá | quebecuá (C) | quebecuá (C) |
| folcloro | (B) | folclor (S) |

Figura 4. Errores léxicos en las transcripciones

Tanto el rehablador como el programa de dictado se enfrentan a la variación a nivel léxico y fonológico. Para que el programa de dictado pueda representar correctamente, en cualquiera de las transcripciones, una palabra pronunciada en la grabación original ya sea en español, incluyendo sus variantes, como en otro idioma (también las originadas por contacto), dicha palabra debe figurar en el léxico del programa. Si la variante *quebecuá* fue reconocida correctamente tanto en la transcripción automática (T1) como en la rehablada (T2) fue porque yo la había añadido al léxico del programa de dictado *Dragon*. Sin embargo, las voces francesas *quartier* y *marchés*, que no formaban parte del léxico de *Dragon*, no podían aparecer en ninguna de las transcripciones; en su lugar, el programa de dictado les atribuyó equivalencias probables, originando errores (S). Por su parte, el rehablador tiende a no repetir una palabra que no forme parte de su propio vocabulario; además, si es experto, tampoco repetirá ninguna que crea que no figura en el diccionario del programa, como veremos más adelante.

La experiencia del rehablador ha permitido marcar el acento prosódico para desambiguar algunos monosílabos (dé, más, tú, él, mí, té, sí, sé), pronombres interrogativos y exclamativos (qué, quién, quiénes, cuál, cuáles, cuánto, cuánta, cuántos, cuántas, cómo), pero, sobre todo, se ha logrado un mejor reconocimiento ante solapamientos, asentimientos y vacilaciones, muy característicos de la conversación. Por ejemplo, en la entrevista E4 el entrevistador asiente doscientas veinticuatro veces. Veamos, en la figura 5, un fragmento de la transcripción de la entrevista E24 y el rendimiento de *Dragon* con la estrategia aplicada por el rehablador cuando el entrevistador asiente. Las líneas en blanco no aparecen en T2, las hemos añadido para ayudar a la lectura.

| **T0** | **T2-L0** |
|---|---|
| mejoramiento de lo que es la calidad de vida. | mejoramiento de lo que es la calidad de vida. |
| [E:] [ASERT] ¡Hum! | [E:] |
| [I:] Porque en… en… en Panamá a pesar de que la moneda, ah… | [I:] Porque en Panamá a pesar de que la moneda |
| [E:] [ASERT] ¡Hum! | |
| [I:] …local es el balboa… | [E:] |
| [E:] [ASERT] ¡Hum! | [I:] Local es el balboa, |
| [I:] …a raíz del… del canal… | [E:] |
| [E:] [ASERT] ¡Hum! | [I:] A raíz del, del canal |
| [I:] …el balboa, que solamente se… se da en moneda… | [E:] |
| [E:] [ASERT] ¡Hum! | [I:] El balboa que solamente se moneda |
| [I:] …no se… no existe en papel, tiene el valor… el valor equivalente monetario al dólar americano. | [E:] |
| | [I:] No existe en papel, tiene el valor el valor equivalente monetario al dólar americano. |

Figura 5. Fragmento 1 (E4): estrategias de rehablado

Cada vez que el entrevistador asiente, el rehablador ha optado por utilizar un atajo verbal que le sirve para que el programa de dictado identifique al entrevistador seguido de un atajo verbal que le sirve para que el programa de dictado identifique al informante. El texto resultante (T2) aparece con la etiqueta [E:] seguida de tabulador, nueva línea, [I:], tabulador. Antes de revisar la transcripción (T2), una búsqueda rápida le permite añadir [ASERT] ¡Hum! en un par de clics. A modo de comparación, este operador de asentimiento aparece ciento dieciocho veces en la entrevista E5, ciento setenta y ocho en la E26 y doscientas cincuenta y cinco en la E27.

En la figura 6 puede verse un ejemplo de las vacilaciones y solapamientos que caracterizan la comunicación oral. De nuevo, las líneas en blanco no aparecen en T2, las hemos añadido para ayudar a la lectura.

| T0 | T2-L0 |
|---|---|
| [I:] Eh… es… es que uno de sus… eh… bisabuelos… | [I:] Es que uno de sus… bisabuelos |
| [E:] [ASERT] Ajá. | |
| [I:] …había sido un gran… eh gram-... eh… gramático… | [I:] había sido un gran gran dramático. |
| [E:] [ASERT] Ajá. | |
| [I:] …después literato. Eh… la familia tenía como un lado de musical y otro lado por… eh… por la lengua. | [I:] Después literato. La familia tenía acumulado de musical y otro lado por por la lengua. |
| [E:] [ASERT] ¡Hum! | |
| [I:] ¿Ya? | [E:] |
| [E:] [ASERT] ¡Hum! | [I:] Ya |
| [I:] Y… y fue muy interesante, eh… porque… | [E:] |
| eh, eh… en tu país ocurre una cosa que en mi país no pasa y es que eh… eh… ustedes son más abiertos que nosotros. | [I:] Y y fue muy interesante porque una cosa que mi país no pasa es que ustedes son más abiertos que nosotros. |
| [E:] Sí. | [E:] |
| [I:] Ya. O sea, en Chile… | [I:] Ya |
| [E:] [SOLAPAMIENTO] No he estado en Chile, pero digamos que sí… sí… sí es… [ASERT] mhm… | [E:] Y agregamos que sí sí |
| [I:] [CHASQUIDO] Y cuando tú llegas a un sector social determinado es casi impenetrable. | [I:] [CHASQUIDO] y cuando tú llegas a un sector social determinado es casi impenetrable. |
| [E:] [ASERT] ¡Hum! | [E:] |

Figura 6. Fragmento 2 (E4): estrategias de rehablado

Ante solapamientos y vacilaciones, el rehablador L0 solo repite lo que está seguro de que le permitirá obtener una suerte de documento esqueleto que podrá editar posteriormente rellenando huecos, en lugar de tener que borrar y reescribir.

## 6. CONCLUSIONES Y PERSPECTIVAS FUTURAS

A la hora de crear un corpus, la transcripción de grabaciones orales puede convertirse en uno de los aspectos más largos del proceso. Para saber qué método al alcance de todos ofrece más ventajas, he comparado la mecanografía, el reconocimiento automático del habla y el rehablado *off-line*. Las estrategias de rehablado *off-line* me han permitido el uso de un programa automático de dictado en su estado actual como herramienta para potenciar la transcripción de las entrevistas del COLEM en menos tiempo y con menos errores.

Con las transcripciones rehabladas hemos reducido la parcialidad y la carga interpretativa. No obstante, sea cual sea el método empleado, el trabajo de transcripción, intenso y difícil, origina muchísima variación (Roulston et al. 2003: 657). Por una parte, el traslado intersemiótico que supone la transcripción de la lengua hablada, al ceñirla dentro de las convenciones escritas, nos obliga a seleccionar qué datos somos capaces de trasladar. Por otra, por muy detallada y completa que sea una transcripción, esta siempre habrá

pasado por el filtro de la interpretación del transcriptor (Kvale 1996) a la vez que, de manera paradójica, el resultado presenta varios grados de heterogeneidad.

La cuestión primordial es definir lo que se considera una transcripción correcta; en su ausencia, los datos que obtendremos serán muy dispares. Esto explica que, por un lado, hayamos encontrado investigadores que afirman que quien teclee rápido puede transcribir una entrevista de una hora entre tres y seis horas (Punch y Oancea 2014; Walford 2001); mientras que, por el otro, hay investigadores que mantienen que la tarea de mecanografía podría estar más cerca de sesenta horas por cada hora registrada de datos, dependiendo del tipo de transcripción que se realice (Evers 2011). El nivel de transcripción, esto es, la cantidad de contenido transcrito es la que está al origen de tales diferencias. El protocolo de transcripción contrarresta la ineludible parcialidad del proceso, minimiza la variabilidad de los resultados, delimita qué debe reflejar cada transcripción y cómo habrá de hacerlo, a la vez que sistematiza la carga interpretativa de signos semióticos propios de una conversación (Cook 1995).

Hay que mencionar la dificultad añadida que supone el hecho de que la representación escrita del discurso oral no es biunívoca. Para ilustrar la complejidad de la tarea, a modo de ejemplo, retomaremos la representación escrita de las unidades léxicas que sirven para construir la interacción. Estos retrocanalizadores lingüísticos (Bravo Cladera, 2009) o *back channel responses* (Yngve, 1970), ratifican la receptividad y atención continuada del interlocutor. Tottie (2014) los llama *vocalizations* y los define como señales o muestras de respuesta que pueden tener varios significados, como asegurarle a nuestro interlocutor que lo estamos siguiendo sin que haya necesidad de retomar el turno de palabra; también pueden tener varias funciones como expresar vacilación, duda, reparo, asombro, asentimiento, negación, etc. En el discurso se caracterizan por tener muchas variantes fonológicas y suelen acompañarse de respuestas no verbales (mover la cabeza de delante a atrás o de un lado a otro, encogerse de hombros), las cuales, en la transcripción, se pierden. Tampoco todos tienen una ortografía normalizada, además, están sometidos al fenómeno de la variación ya que en varias lenguas poseen similitudes sonoras con sus propias variantes entonativas y de contenido semántico. Estos retrocanalizadores lingüísticos aparecen con tal frecuencia en las entrevistas que han ocasionado mucha inversión de tiempo por parte de transcriptores y revisores. En los corpus, la variación en su representación "dificulta una consulta comparable y, en consecuencia, la descripción de sus usos" (Solís García y León Gómez, 2018: 332). Para nuestra investigación se ha propuesto una etiqueta única acompañada de tres codificaciones: la interjección hum, que es la representación más generalizada (Bravo Cladera, 2009), *mhm.,* cuando el alargamiento es más consonántico y *ehm* cuando el contenido fónico es más vocálico. La etiqueta debe facilitar la automatización del lenguaje y el rehablado, así como la transcripción y la revisión, dejando el corpus abierto a futuros estudios de otra índole: lingüísticos, de análisis del discurso, de pragmática.

A modo de reflexión final, señalamos la apremiante necesidad de encontrar un sistema que integre todas las herramientas para la elaboración de materiales de estudio de corpus: transcripción, revisión, etiquetado y análisis de datos. Entendemos que la clave está en una colaboración eficaz humano-computadora. Centrándonos en nuestro objeto de estudio, la transcripción, los creadores de interfaces contemplan la corrección de errores del RAH desde la perspectiva de la transcripción asistida por ordenador (Revuelta-Martínez et al., 2012). Esto es, permitir la intervención del humano en el momento en que el programa está produciendo el reconocimiento automático para corregir, ya sea con teclado, ratón, voz, gestos o atajos, en tiempo real, a imagen de los modelos predictivos usados en traducción

asistida por ordenador (de Souza et al., 2015). Dentro de este enfoque, el rehablado *off-line* debería poder usarse para editar los segmentos que contienen más errores sobre la marcha (Sperber et al., 2017; Sperber et al., 2016). En definitiva, gracias a la combinación de herramientas que hasta ahora se han empleado de forma aislada se lograrán mejores resultados en menor tiempo y la metodología de transcripción dará un gran paso adelante.

## AGRADECIMIENTOS

Nuestro agradecimiento a los participantes en el taller de rehablado *off-line*, al equipo del CRIM, especialmente a Simon Desrochers, por la colaboración con nuestro proyecto para el análisis de datos y a Enrique Pato por la lectura del manuscrito.

## REFERENCIAS BIBLIOGRÁFICAS

AA.VV. (2008). Diccionario de términos clave de ELE. Diccionario de términos clave de ELE. https://cvc.cervantes.es. [Consultado el 29 de junio de 2020].

ACR/CAB (2012). *Normes universelles du sous-titrage codé à l'intention des télédiffuseurs canadiens de langue française*. Document révisé en fonction de CRTC 2011-741 et CRTC 2011–741-1. http://www.cab-acr.ca. [Consultado el 25 de junio de 2020].

Adolphs, S., Knight, D. (2010). Building a spoken corpus. What are the basics? In: A. O'keeffe, M. Mccarthy (eds.), *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 38-52.

AENOR (2006). *Norma Española UNE-EN 15038. Servicios de traducción. Requisitos para la prestación del servicio*. Madrid: AENOR.

Ainsworth, W. A. (1988). *Speech recognition by machine*. London, UK: Institution of Electrical Engineers.

Bernabé Caro, R., Eugeni, C., Masia, V., Oncins, E. (2019). Bridging the gap between training and profession in real-time intralingual subtitling. Media for All 8, Stockholms universitet, Stockholm, 17-19 de junio de 2019.

Blackley, S. V., Huynh, J., Wang, L., Korach, Z., Zhou, L. (2019). Speech recognition for clinical documentation from 1990 to 2018: a systematic review. *Journal of the american medical informatics association*, 26(4). 324–338. doi: 10.1093/jamia/ocy179.

Bravo Cladera, N. (2009). La retrocanalización como una realización de la interacción: algunos usos de mm y mhm en español. In: *Actas del II Congreso de Hispanistas y Lusitanistas Nórdicos: Actas do II Congresso de Hispanistas e Lusitanistas Nórdicos*, Estocolmo, 25-27 de octubre de 2007: Acta Universitatis Stockholmiensis, Romanica Stockholmiensia 26, 25-42.

Brinton, B., Fujiki, M., Winkler, E., Loeb, D. F. (1986). Responses to requests for clarification in linguistically normal and language-impaired children. *Journal of Speech & Hearing Disorders*, 51(4). 370-378. doi: 10.1044/jshd.5104.370.

Brooks, C. (2010). Embodied Transcription: A Creative Method for Using Voice-Recognition Software. *Qualitative Report*, 15(5), pp. 1227-1241.

Brousseau, J., Beaumont, J.-F., Boulianne, G., Cardinal, P., Chapdelaine, C., Comeau, M., Osterrath, F., Ouellet, P. (2003). Automated closed-captioning of live TV broadcast news in French. In: *Eurospeech 2003 Proceedings: 8th European Conference on Speech Communication and Technology*, INTERSPEECH 2003, Geneva, CICG, 1-4 de septiembre de 2003. Bonn: ISCA, 1245-1248.

Chiu, C.-C., Tripathi, A., Chou, K., Co, C., Jaitly, N., Jaunzeikare, D., Kannan, A., Nguyen, P., Sak, H., Sankar, A., Tansuwan, J., Wan, N., Wu, Y., Xuedong, Z. (2018). Speech recognition for medical conversations. In: *19th International Conference on Acoustics, Speech and Signal*

*Processing (ICASSP)*, INTERSPEECH 2018, Hyderabadm, International Convention Centre, 2-6 de septiembre de 2018: ISCA, 2972-2976.

Cook, G. (1995). Theoretical issues: transcribing the untranscribable. In: G. Leech, G. Myers, J. Thomas (eds.), *Spoken English on computer: Transcription, mark-up and application*. Harlow, UK: Longman, 35-53.

CRTC, C. d. l. r. e. d. t. c. (2007). *Nouvelle politique de sous-titrage codé pour malentendants*. 2007-54. Ottawa: Gouvernement du Canada. https://crtc.gc.ca. [Consultado el 20 de junio de 2020].

D'Arcangelo, R., Cellini, F. (2013). Metodi e tempi di una verbalizzazione-prove tecniche. In: C. Eugeni, L. Zambelli (eds.). *Specializzazione on-line. Numero monografico sul Respeaking*(Maggio 2013), pp. 81-95. http://www.accademia-aliprandi.it.

Damper, R. I., Lambourne, A. D., Guy, D. P. (1985). Speech input as an adjunct to keyboard entry in television subtitling. In: B. Shackel (ed.), *Human-Computer Interaction-INTERACT'84*. Amsterdam: Elsevier, 203-208.

Davidson, C. (2009). Transcription: Imperatives for Qualitative Research. *International Journal of Qualitative Methods*, 8(2). 35-52. doi: 10.1177/160940690900800206.

de Souza, J. G., Negri, M., Ricci, E., Turchi, M. (2015). Online multitask learning for machine translation quality estimation. In: C. Zong, M. Strube (eds.), *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (Vol. 1: Long papers), IJCNLP 2015, Beijing, China National Convention Center, 26-31 de julio de 2015. Red Hook, NY: ACL, 219-228.

Dudley, H. (1958). Phonetic Pattern Recognition Vocoder for Narrow-Band Speech Transmission. *The Journal of the Acoustical Society of America*, 30(8), pp. 733-739.

Durand, J. (2017). Corpus Phonology. *Oxford Research Encyclopedia of Linguistics*.

Edwards, E., Salloum, W., Finley, G. P., Fone, J., Cardiff, G., Miller, M., Suendermann-Oeft, D. (2017). Medical Speech Recognition: Reaching Parity with Humans. In: A. Karpov, R. Potapova, I. Mporas (eds.), *Speech and Computer. 19th International Conference, SPECOM 2017 Hatfield, UK, September 12–16, 2017 Proceedings* (Vol. 10458). Cham, CH: Springer International Publishing, 512-524. doi: 10.1007/978-3-319-66429-3 51.

El español en Montreal: COLEM (2006). Corpus: Pato, Enrique (dir.). https://esp-montreal.jimdo.com. [Consultado el 17 de abril de 2020].

Elvira-García, W., Roseano, P., Fernández Planas, A., Martínez Celdrán, E. (2015). Una herramienta para la transcripción prosódica automática con etiquetas Sp_ToBI en Praat. In: A. Cabedo Nebot (ed.), *Perspectivas actuales en el análisis fónico del habla: tradición y avances en la fonética experimental*. València: Universitat de València, 455-464.

Evers, J. C. (2011). From the past into the future. How technological developments change our ways of data collection, transcription and analysis. *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, 12(1). Art. 38.

FCC (2014). *Report and Order, Declaratory Ruling, and Further notice of proposed rulemaking*. FCC-14 12A1. Federal Communications Commission. https://www.fcc.gov/. [Consultado el 25 de junio de 2020].

Ferber, R. (1991). Slip of the tongue or slip of the ear? On the perception and transcription of naturalistic slips of the tongue. *Journal of Psycholinguistic Research*, 20(2), pp. 105-122.

Fogg, T., Wightman, C. W. (2000). Improving Transcription of Qualitative Research Interviews with Speech Recognition Technology. Creating Knowledge in the 21st Century: Insights From Multiple Perspectives (American Educational Research Association Annual Meeting), New Orleans, LA, 24-26 de abril de 2000.

Gadet, F., Ludwig, R., Mondada, L., Pfänder, S., Simon, A. C. (2012). Un grand corpus de français parlé: le CIEL-F. *Revue française de linguistique appliquée*, 17(1). 39-54.

Hawkins, W. R., Robinson, R. N. (1979). The development and use of an electric keyboard for television subtitling by Palantype. *International Journal of Man-Machine Studies*, 11(6), pp. 701-710.

Hodgson, T., Coiera, E. (2016). Risks and benefits of speech recognition for clinical documentation: a systematic review. *Journal of the american medical informatics association*, 23(e1), pp. 169-179. doi: 10.1093/jamia/ocv152.

Ivarsson, J. (1992). *Subtitling for the media: A handbook of an art*. Stockholm: Transedit.

Johnson, B. E. (2011). The speed and accuracy of voice recognition software-assisted transcription versus the listen-and-type method: a research note. *Qualitative Research*, 11(1), pp. 91-97. doi: 10.1177/1468794110385966.

Johnson, M., Lapkin, S., Long, V., Sanchez, P., Suominen, H., Basilakis, J., Dawson, L. (2014). A systematic review of speech recognition technology in health care. *BMC medical informatics and decision making*, 14(1). doi: 10.1186/1472-6947-14-94.

Kreuz, R. J., Riordan, M. A. (2018). The art of transcription: Systems and methodological issues. In: A. H. Jucker, K. P. Schneider, W. Bublitz (eds.), *Methods in Pragmatics* (Vol. 10). Berlin-Boston: The Gruiter Mouton, 1634.

Kvale, S. (1996). *InterViews: An introduction to qualitative research interviewing*. Thousand Oaks, CA: Sage.

Lambourne, A. D. (2007). Re-speaking the truth. *International Broadcast Engineer*, IBE JULY/AUGUST 2007: 44-46.

Lapadat, J. C. (2000). Problematizing transcription: Purpose, paradigm and quality. *International Journal of Social Research Methodology*, 3(3), pp. 203-219. doi: 10.1080/13645570050083698.

Li Deng, Yang Liu, ed. (2018). *Deep learning in natural language processing*. Singapore: Springer Singapore. doi: 10.1007/978-981-10-5209-5.

Lindsay, J., O'Connell, D. C. (1995). How do transcribers deal with audio recordings of spoken discourse? *Journal of Psycholinguistic Research*, 24(2), pp. 101-115.

Lu, X., Li, S., Fujimoto, M. (2020). Automatic speech recognition. In: Y. Kidawara, E. Sumita, H. Kawai (eds.), *Speech-to-Speech Translation*. Singapore: Springer Singapore, 21-38. doi: 10.1007/978-981-15-0595-9_2.

MacLean, L. M., Meyer, M., Estable, A. (2004). Improving accuracy of transcripts in qualitative research. *Qualitative health research*, 14(1), pp. 113-123. doi: 10.1177/1049732303259804.

Manrique Fuero, F. (2016). Innovación en técnicas de estenotipia para subtitulado en directo. In: *VIII Congreso de Accesibilidad a los Medios Audiovisuales para Personas con Discapacidad*, AMADIS'16, Toledo, Museo del Ejército del Alcázar de Toledo, 27-28 de octubre de 2016. Madrid: Real Patronato sobre Discapacidad, 44-58.

Mariani, J., ed. (2002). *Analyse, synthèse et codage de la parole. Traitement automatique du langage parlé 1* (vol 1). Paris: Hermès Science Publications.

Markle, D. T., West, R. E., Rich, P. J. (2011). Beyond transcription: technology, change, and refinement of method. *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, 12(3). Art. 21.

Matamala, A., Romero-Fresco, P., Daniluk, L. (2017). The use of respeaking for the transcription of non-fictional genres: An exploratory study. *inTRAlinea: Online Translation Journal,* 19. http://www.intralinea.org.

Matheson, J. L. (2007). The voice transcription technique: Use of voice recognition software to transcribe digital interview data in qualitative research. *The Qualitative Report*, 12(4), pp. 547-560.

McCoy, E., Shumway, R. (1979). Real-time captioning--promise for the future. *American Annals of the Deaf*, 124(5), pp. 681-690.

Moores, Z. (2016). Subtitling live events through respeaking-Increasing accessibility for all. Unlimited! International Symposium on Accessible Live Events, University of Antwerp, Antwerp, 29 de abril de 2016.

Moro Vallina, M. (2010). *Aplicaciones ofimáticas*. Madrid: Editorial Paraninfo.

Niemants, N. (2018). Des enregistrements aux corpus : transcription et extraction de données d'interprétation en milieu médical. *Meta*, 63(3), pp. 665-694.

Núñez Hidalgo, J., Ramos Villajos, E. (2010). Doscientos años de taquigrafía parlamentaria: de las Cortes de Cádiz a nuestros días (1810-2010). *Revista de las Cortes Generales*, (80), pp. 145-179.

NVRA, N. V. R. A. (2008). The Horace Webb history. https://nvra.org. [Consultado el 2 de febrero de 2018].

Ochs, E. (1979). Transcription as theory. In: Elinor Ochs, Bambi Schieffelin (eds.), *Developmental pragmatics*. New York, NY: Academic Press, 43-72.

ONU, N. U. (1994). *Normas Uniformes sobre la igualdad de oportunidades para las personas con discapacidad*. Resolución 48/96. https://www.un.org. [Consultado el 20 de junio de 2020].

Park, J., Zeanah, A. E. (2005). An evaluation of voice recognition software for use in interview-based research: A research note. *Qualitative Research*, 5(2), pp. 245-251.

Pato, E. (2017). La realidad lingüística en Canadá y la situación del español en Montreal. *Oltreoceano*, 13: 27-37. doi: 10.1400/253045.

Pollard, S. E., Neri, P. M., Wilcox, A. R., Volk, L. A., Williams, D. H., Schiff, G. D., Ramelson, H. Z., Bates, D. W. (2013). How physicians document outpatient visit notes in an electronic health record. *International Journal of Medical Informatics*, 82(1). 39-46.

Punch, K. F., Oancea, A. (2014). *Introduction to research methods in education* (2nd edition). SAGE Publishing. www.sagepublications.com. [Consultado el 23 de junio de 2020].

Ravanelli, M., Brakel, P., Omologo, M., Bengio, Y. (2018). Light gated recurrent units for speech recognition. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(2), pp. 92-102. doi: 10.1109/TETCI.2017.2762739.

Revuelta-Martínez, A., Rodríguez, L., García-Varea, I. (2012). A computer assisted speech transcription system. In: F. Segond (ed.), *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL 2012, Avignon, 23-27 de abril de 2012. Stroudsburg, PA: Association for Computational Linguistics, 41-45. doi: 10.3115/v1/P15-1.

Revuelta Domínguez, F. I., Sánchez Gómez, M. C. (2005). El proceso de transcripción en el marco de la metodología de investigación cualitativa actual. *Enseñanza & Teaching*, 23: 367-386.

Rodero Antón, E. (2016). Influence of speech rate and information density on recognition: The moderate dynamic mechanism. *Media Psychology*, 19(2), pp. 224-242. Consultado el 2016/04/02. doi: 10.1080/15213269.2014.1002942.

Romero-Fresco, P. (2011). *Subtitling through speech recognition: respeaking* (1a). Manchester, UK; Kinderhook, NY: St. Jerome Publishing.

Romero-Fresco, P. (2012). Respeaking in translator training curricula, present and future prospects. *The Interpreter and Translator Trainer*, 6(1), pp. 91-112. doi: 10.1080/13556509.2012.10798831.

Roulston, K., DeMarrais, K., Lewis, J. B. (2003). Learning to interview in the social sciences. *Qualitative inquiry*, 9(4), pp. 643-668.

Rufino Morales, M. (2020). El rehablado *off-line* para optimizar la transcripción de corpus orales en español. In: S. Martínez Martínez (ed.), *"Nuevas tendencias en Traducción e Interpretación"*. Granada: Comares, 17-27.

Saon, G., Kurata, G., Sercu, T., Audhkhasi, K., Thomas, S., Dimitriadis, D., Cui, X., Ramabhadran, B., Picheny, M., Lim, L.-L. (2017). English conversational telephone speech recognition by humans and machines. *arXiv preprint arXiv:1703.02136*.

Solís García, I., León Gómez, M. (2018). Hum, operador de la interacción en español: valor y usos. *Círculo de Lingüística Aplicada a la Comunicación*, 74: 323-352.

Sperber, M., Neubig, G., Nakamura, S., Waibel, A. (2016). Optimizing Computer-Assisted Transcription Quality with Iterative User Interfaces. In: N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (eds.), *Proceedings of the 10th International Conference on Language Resources and Evaluation Conference (LREC 2016)*, Portoroz, Sl, Grand Hotel Bernardin Conference Center, 23-28 de mayo de 2016: European Language Resources Association (ELRA), 1986-1992.

Sperber, M., Neubig, G., Niehues, J., Nakamura, S., Waibel, A. (2017). Transcribing against time. *Speech communication*, 93: 20-30.

Tanton, N. E. (1979). UK Teletext-Evolution and Potential. *IEEE Transactions on Consumer Electronics*, CE-25(3), pp. 246-250. doi: 10.1109/TCE.1979.273220.

Tatham, M., Morton, K. (2005). *Developments in speech synthesis*. Chichester, WS: John Wiley & Sons.

Tilley, S. A. (2003). "Challenging" research practices: Turning a critical lens on the work of transcription. *Qualitative inquiry*, 9(5), pp. 750-773.

Tottie, G. (2014). What does uh-(h) uh mean: American English vocalizations and the Swedish learner. *Functions of Language*, 21(1), pp. 6-29.

Utray Delgado, F., García Castillejo, Á., Puente Rodríguez, L., Carrero Lea, J. M., Pajares, J. L., González, Y., Ruiz-Mezcua, B., Sánchez Pena, J. M. (2015). Informe de seguimiento del subtitulado y la audiodescripción en la TDT 2014, *Colección Inclusión y Diversidad*. 15. Madrid: Grupo Editorial Cinca.

Walford, G. (2001). *Doing qualitative educational research. A personal guide to research process*. London, UK; New York, NY: Bloomsbury Publishing.

Winata, G. I., Cahyawijaya, S., Liu, Z., Lin, Z., Madotto, A., Xu, P., Fung, P. (2020). Learning fast adaptation on cross-accented speech recognition. *arXiv preprint arXiv:2003.01901*.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., Zweig, G. (2016). Achieving human parity in conversational speech recognition. *arXiv:1610.05256*.

Yadav, D., Gupta, M., Chetlur, M., Singh, P. (2018). Automatic annotation of voice forum content for rural users and evaluation of relevance. In: *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*, Menlo Park, CA, 20-22 de junio de 2018. doi: 10.1145/3209811.3209875.

Yngve, V. (1970). On getting a word in edgewise. In: *Papers from the sixth regional meeting, Chicago Linguistic Society*, Chicago, IL, 16-18 de abril de 1970: Chicago Linguistic Society, 567-577.

Zhou, L., Blackley, S. V., Kowalski, L., Doan, R., Acker, W. W., Landman, A. B., Kontrient, E., Mack, D., Meteer, M., Bates, D. W., Goss, F. R. (2018). Analysis of Errors in Dictated Clinical Documents Assisted by Speech Recognition Software and Professional Transcriptionists. *JAMA Network Open*, 1(3). e180530-e180543. doi: 10.1001/jamanetworkopen.2018.0530.

Zweigenbaum, P., Gauvain, J.-L., Braffort, A., Filhol, M., Ghannay, S., Grouin, C., Hamon, T., Illouz, G., Lavergne, T., Ligozat, A.-L. (2020). Le bulletin n°107 – janvier 2020 – IA et Technologies du Langage Humain. Grenoble: Association française pour l'Intelligence Artificielle (AFIA). https://afia.asso.fr.