VI Jornadas de Investigación en Humanidades

Homenaje a Cecilia Borel

Departamento de Humanidades

Universidad Nacional del Sur

30 de noviembre al 2 de diciembre de 2015

EDITORIAL
DE LA UNIVERSIDAD
NACIONAL DEL SUR

VI Jornadas de Investigación en Humanidades: homenaje a Cecilia Borel / Daiana Agesta... [et al.]; editado por Omar Chauvié ... [et al.]. - 1a ed. - Bahía Blanca: Editorial de la Universidad Nacional del Sur. Ediuns, 2019.
Libro digital, PDF
Archivo Digital: descarga y online

ISBN 978-987-655-222-6

1. Humanidades. 2. Investigación. I. Agesta, Daiana II. Chauvié, Omar, ed.
CDD 300.72

Editorial de la Universidad Nacional del Sur |
Santiago del Estero 639 | B8000HZK Bahía Blanca | Argentina
www.ediuns.com.ar | ediuns@uns.edu.ar
Facebook: EdiUNS | Twitter: EditorialUNS

Libro
Universitario
Argentino

Diseño interior: Alejandro Banegas
Diseño de tapa: Fabián Luzi

Queda hecho el depósito que establece la Ley n.° 11723.
Bahía Blanca, Argentina, julio de 2019.

VI Jornadas de Investigación en Humanidades "Homenaje a Cecilia Borel"
Departamento de Humanidades - Universidad Nacional del Sur
30 de noviembre al 2 de diciembre de 2015

Coordinación
Lic. Laura Orsi

Declaradas de Interés Municipal por la ciudad de Bahía Blanca.
Declaradas de Interés Educativo por la provincia de Buenos Aires en la sesión del 4 de septiembre de 2015 Resolución n.º 1665/2015-, Expediente n.° 5801361392/15

Autoridades
Universidad Nacional del Sur

Rector: Dr. Mario Ricardo Sabbatini
Vicerrectora: Mg. Claudia Patricia Legnini
Secretario General de Ciencia y Tecnología: Dr. Sergio Vera
Departamento de Humanidades
Directora Decana: Lic. Silvia T. Álvarez
Vicedecana: Lic. Laura Rodríguez
Secretario Académico: Dr. Leandro Di Gresia
Secretaria de Investigación, Posgrado y Formación Continua: Lic. Laura Orsi
Secretario de Extensión y Relaciones Institucionales: Lic. Diego Poggiese

Comisión Organizadora

Srta. Daiana Agesta
Dra. Marcela Aguirrezabala
Dr. Sebastián Alioto
Lic. Carolina Baudriz
Lic. Clarisa Borgani
Prof. Lucas Brodersen
Lic. Gonzalo Cabezas
Dra. Rebeca Canclini
Lic. Norma Crotti
Srta. Victoria De Angelis

Lic. Mabel Díaz
Dra. Marta Domínguez
Srta. M. Bernarda Fernández Vita
Srta. Ana Julieta García
Srta. Florencia Garrido Larreguy
Dra. M. Mercedes González Coll
Mg. Laura Iriarte
Sr. Lucio Emmanuel Martin
Mg. Virginia Martin
Esp. Andrea Montano
Lic. Lorena Montero
Psic. M. Andrea Negrete
Srta. M. Belén Randazzo
Dra. Diana Ribas
Srta. Valentina Riganti
Sr. Esteban Sánchez
Mg. Viviana Sassi
Lic. José Pablo Schmidt
Dra. Marcela Tejerina
Dra. Sandra Uicich
Prof. Denise Vargas

Comisión Académica

Dr. Sandro Abate (Universidad Nacional del Sur – CONICET)
Dra. Marcela Aguirrezabala (Universidad Nacional del Sur)
Dra. Ana María Amar Sánchez (Universidad de California, Irvine)
Dra. Marta Alesso (Universidad Nacional de La Pampa)
Dra. Adriana María Arpini (Universidad Nacional de Cuyo)
Dr. Marcelo Auday (Universidad Nacional del Sur)
Dr. Eduardo Azcuy Ameghino (Universidad de Buenos Aires – CONICET)
Dr. Fernando Bahr (Universidad Nacional del Litoral – CONICET)
Dra. M. Cecilia Barelli (Universidad Nacional del Sur – CONICET)
Dr. Raúl Bernal Meza (Universidad del Centro de la Provincia de Bs. As.)
Dr. Hugo Biagini (Universidad Nacional de La Plata – CONICET)
Dr. Lincoln Bizzozero (Universidad de La República, Uruguay)
Dra. Mercedes Isabel Blanco (Universidad Nacional del Sur)
Dr. Gustavo Bodanza (Universidad Nacional del Sur – CONICET)
Dra. Nidia Burgos (Universidad Nacional del Sur)
Dr. Roberto Bustos Cara (Universidad Nacional del Sur)
Dra. Mabel Cernadas (Universidad Nacional del Sur – CONICET)
Dra. Laura Cristina del Valle (Universidad Nacional del Sur)
Dr. Eduardo Devés (Universidad de Santiago de Chile)
Dra. Marta Domínguez (Universidad Nacional del Sur)
Dr. Oscar Esquisabel (Universidad Nacional de La Plata – CONICET)

Dra. Claudia Fernández (Universidad Nacional de La Plata – CONICET)
Dra. Ana Fernández Garay (Universidad Nacional de La Pampa – CONICET)
Dra. Estela Fernández Nadal (Universidad Nacional de Cuyo – CONICET)
Dr. Rubén Florio (Universidad Nacional del Sur)
Dra. Lidia Gambon (Universidad Nacional del Sur)
Dr. Ricardo García (Universidad Nacional del Sur)
Dra. Viviana Gastaldi (Universidad Nacional del Sur)
Dr. Alberto Giordano (Universidad Nacional de Rosario)
Dra. Graciela Hernández (Universidad Nacional del Sur – CONICET)
Dra. Yolanda Hipperdinger (Universidad Nacional del Sur – CONICET)
Dra. Silvina Jensen (Universidad Nacional del Sur – CONICET)
Dr. Juan Francisco Jimenez (Universidad Nacional del Sur)
Dra. María Mercedes González Coll (Universidad Nacional del Sur)
Dra. María Luisa La Fico Guzzo (Universidad Nacional del Sur)
Dr. Javier Legris (Universidad de Buenos Aires – CONICET)
Dra. Celina Lértora (Universidad del Salvador – CONICET)
Dr. Fernando Lizárraga (Universidad Nacional del Comahue - CONICET)
Dra. Elisa Lucarelli (Universidad de Buenos Aires)
Mg. Ana María Malet (Universidad Nacional del Sur)
Prof. Raúl Mandrini (Universidad Nacional del Centro de la Provincia de Bs. As.)
Dra. Stella Maris Martini (Universidad de Buenos Aires)
Dr. Raúl Menghini (Universidad Nacional del Sur)
Dra. Elda Monetti (Universidad Nacional del Sur)
Dr. Rodrigo Moro (Universidad Nacional del Sur – CONICET)
Dra. Lidia Nacuzzi (Universidad de Buenos Aires – CONICET)
Dr. Ricardo Pasolini (Universidad Nacional del Centro de la Provincia de Bs. As.)
Dr. Sergio Pastormerlo (Universidad Nacional de La Plata)
Dra. Dina Picotti (Universidad de Buenos Aires – CONICET)
Dr. Luis Porta (Universidad Nacional de Mar del Plata – CONICET)
Dra. M. Alejandra Pupio (Universidad Nacional del Sur)
Dra. Alicia Ramadori (Universidad Nacional del Sur)
Dra. Silvia Ratto (Universidad de Buenos Aires)
Dra. Diana Ribas (Universidad Nacional del Sur)
Dra. Elizabeth Rigatuso (Universidad Nacional del Sur – CONICET)
Lic. Adriana Rodríguez (Universidad Nacional del Sur)
Dr. Hernán Silva (Universidad Nacional del Sur – CONICET)
Dra. Marcela Tejerina (Universidad Nacional del Sur)
Dr. Fernando Tohmé (Universidad Nacional del Sur – CONICET)
Dra. Fabiana Tolcachier (Universidad Nacional del Sur)
Dra. Patricia Vallejos (Universidad Nacional del Sur – CONICET)
Dra. Irene Vasilachis (CEIL – CONICET)
Dra. María Celia Vázquez (Universidad Nacional del Sur)
Dr. Daniel Villar (Universidad Nacional del Sur)
Dr. Emilio Zaina (Universidad Nacional del Sur)
Dra. Ana María Zubieta (Universidad de Buenos Aires – CONICET)

Lucía **Cantamutto**
Lorena M. A. **de- Matteis**
Gimena **Del Río**
Federico **Gobato**
Nora **González**
Mónica **Ricca**
(Editores)

Humanidades Digitales y prácticas de enseñanza de la lengua

Volumen 16

Índice

Reflexiones sobre una colaboración incipiente: herramientas informáticas para una investigación en lingüística aplicada[1]

Lorena M. A. de- Matteis
Departamento de Humanidades - Universidad Nacional del Sur - CONICET
lmatteis@uns.edu.ar

Leonardo J. D. de- Matteis
Departamento de Ciencias e Ingeniería de la Computación – Universidad Nacional del Sur
ldm@cs.uns.edu.ar

1. Introducción

Si bien está extendido en el ámbito de las ciencias humanas el consenso sobre la utilidad que las herramientas informáticas pueden tener al optimizar la labor de los investigadores en tanto "operarios" para facilitar su trabajo como "analistas" (Stulic-Etchevers y Rouissi, 2009; Villayandre Llamazares, 2010), y aunque ya es sabido que en un mundo digital se redefinen los objetos de estudio y las prácticas a través de las cuales se los interroga, también resulta cierto que, en muchos casos, los proyectos de investigación final de carrera o incluso de posgrado no suelen incluir el empleo de programas especializados en las humanidades[2]. Excepciones aparte, las encuestas de sondeo realizadas en 2012 en el marco de los preparativos para un curso introductorio sobre aplicaciones informáticas en ciencias humanas en el marco del Departamento de Humanidades de la Universidad Nacional del Sur (UNS) sugerían *a)* un predominio absoluto de uso de herramientas ofimáticas (de carácter autodidacta); *b)* un interés por incorporar nuevas herramientas, aunque condicionado por la falta de oportunidades para el aprendizaje sobre las posibilidades existentes.

En este contexto, el surgimiento de asociaciones nuevas (como la Asociación Argentina de Humanidades Digitales [AAHD]) es indicador de una tendencia de cambio y pone de manifiesto la formación de redes interdisciplinarias nuevas así como la posibilidad de entablar un diálogo fructífero entre humanistas y especialistas en informática que pueden colaborar en la selección, adaptación y, cuando es posible y necesario, en el desarrollo de herramientas apropiadas para la investigación humanística.

Este trabajo intenta reflexionar sobre las instancias iniciales de un esfuerzo de este tipo, entablado entre docentes del Departamento de Humanidades y del de Ciencias e Ingeniería de la Computación de la UNS, con la intención de colaborar en el desarrollo de herramientas para la transcripción y procesamiento de muestras orales recogidas en un proyecto de investigación orientado a la conformación

[1] Trabajo inscripto en el proyecto de investigación "Desarrollo de competencias comunicativas profesionales para la seguridad lingüística: análisis de prácticas de instrucción y desarrollo de herramientas didácticas para el ámbito aeronáutico", P.G.I.24/I220 SGCyT-UNS.

[2] Algunas de las alternativas posibles, al menos en los estudios lingüísticos, son las herramientas cuantitativas como *SPSS*, las cuali-cuantitativas como *WordSmith Tools* y otras más orientadas a lo cualitativo como *ATLAS.ti*.

de un *corpus* multipropósito para un discurso institucional particular como es el del control de tránsito aéreo (ATC). Se revisarán para ello tanto las características de diseño de los *corpus* disponibles para este ámbito socio-técnico —mencionando también distintas herramientas que se pueden utilizar para la transcripción y etiquetado de corpus— como, de manera sucinta, las limitaciones experimentadas hasta la fecha en nuestro trabajo con herramientas ofimáticas tradicionales para la gestión de bases de datos (*Microsoft Access*) y para administrar y procesar *corpus* textuales de pequeño y mediano tamaño (*WordSmith Tools 6.0*). A partir de estas revisiones, el trabajo intentará delinear las principales características de una herramienta especialmente orientada a la construcción de un *corpus* ATC.

2. Antecedentes

La lingüística de *corpus* (Baker, 2010; O'Keeffe y McCarthy, 2010; McEnery y Hardie 2012) señala uno de los principales puntos de contacto entre las humanidades y la informática. Su utilidad y potencialidad provienen de las ventajas evidentes de contar con colecciones de documentos —escritos u orales— que se caracterizan por las propiedades de *accesibilidad, reusabilidad*, *perennidad*, *plasticidad*, *almacenamiento*, *asociabilidad* y *programabilidad,* derivadas a su vez de las posibilidades que su etiquetado y anotación abren a la hora de desarrollar algoritmos para procesar los datos lingüísticos (Stulic-Etchevers y Rouissi, 2009: 119). Además, permiten consolidar avances colectivos a través de una labor que nuclee a investigadores de distintos centros, permitiendo que la comunidad académica contraste, valide y profundice los resultados obtenidos en el análisis de los datos.

2.1 Antecedentes: otros *corpus* digitales para el dominio ATC

Desde la década de 1990 existen *corpus* de habla —es decir, *speech corpus,* o *corpus* de datos orales— creados para la investigación de la comunicación en el dominio aeronáutico (tabla n.° 1). Algunos están orientados a la aviación militar, como el *non-native Military Air Traffic Control* (nnMATC) (Pigeon, *et al.*, 2007) o el HIWIRE (Segura *et al.,* 2007), y otros al ámbito ATC en la aviación civil.

Además de la esfera aeronáutica de la que toman sus datos, los *corpus* ATC pueden clasificarse en dos grupos según los integren datos operacionales —esto es, en el registro de interacciones reales entre pilotos y controladores aéreos— o simulados. Entre los primeros, se cuentan, por ejemplo, el *ATC Complete Corpus*, también conocido como *ATC Corpus* (Godfrey, 1994) y el *Vocalise* (Graglia *et al*., 2005); entre los segundos destaca el *ATCOSIM* (Hofbauer *et al.*, 2008).

Entre las iniciativas más recientes con datos operacionales, además, pueden mencionarse las propuestas menos conocidas del *Spoken Corpus of Radiotelephony Phraseology* (*SCRP*, Pavlinović *et al.*, 2013a y Pavlinović *et al*., 2013b), del *ATCC Speech Corpus* (Šmídl e Ircing, 2014) y la de un *corpus* de inglés aeronáutico como LSP (*ATC-LSP*, Lopez *et al.*, 2013).

La primera observación que puede realizarse es que todos los *corpus* identificados lo son de interacciones ATC en idioma inglés o centran su atención en las situaciones de contacto con esta lengua. Esto se explica tanto por la relevancia del inglés y de su instrucción para la seguridad de la aviación internacional (v. de- Matteis *et al*., 2015), como por las finalidades que persiguen estos *corpus*.

En cuanto a las características de su diseño, puede destacarse la coincidencia en adoptar una transcripción ortográfica de las interacciones. Esta decisión, que también adoptamos en de- Matteis (2009), limita su relevancia para una investigación de tipo fonético-fonológico pero se justifica tanto

por los objetivos prioritarios de estas herramientas como por la dificultad que entraña emprender este tipo de transcripción con grabaciones que muchas veces poseen baja calidad o en las que se detectan interferencias y ruidos que dificultan tal tarea. También en relación con su diseño, puede señalarse que solo algunos adoptan un sistema de etiquetado, seleccionando para la anotación solo algunas características, sin que se anoten fenómenos para todos los niveles de análisis lingüístico.

Por otro lado, mientras que algunas de las propuestas se basan en herramientas informáticas propias (como *WebTransc*), otras se apoyan en *software* estándar para el manejo de *corpus* digitales, como *WordSmith Tools*.

En cuanto a sus fines, pueden señalarse algunas tendencias compartidas: algunos de ellos se proponen modelar la interacción que se produce en el ámbito ATC como paso previo para desarrollar otras herramientas tecnológicas que se orientan, en general, a tres fines específicos: *a)* sistemas de reconocimiento automático de habla (o ASR por sus siglas en inglés), con especial atención al estudio del habla de hablantes no nativos de inglés para mejorar tales sistemas; *b)* tecnologías de ASR que detecten y adviertan sobre desviaciones respecto de la fraseología estandarizada y *c)* herramientas que simulen hablantes en un sistema de diálogo inteligente para reemplazar a los "pseudopilotos" en el entrenamiento de los controladores de tránsito aéreo. Por último, destaca por su intención pedagógica y su interés para nuestra labor, la propuesta de Lopez *et al*. (2013) de constituir un *corpus* de aprendizaje (*learner's corpus*) con la finalidad de asistir en la enseñanza de la radiotelefonía en inglés a controladores franceses.

	ATC Complete Corpus	HIWIRE	nnMATC	VOCALISE	ATCOSIM	SCRP	ATCC	ATC-LSP
Objetivo	ASR	ASR	ASR	s/d	amplio espectro	ASR	ASR	pedagógica
Situación								
a) díada de hablantes	P/C civil	P/C militar	P/C militar	P/C civil	C civil	P/C civil	P/C civil	P/C civil
b) tipo de control	APP, TWR?	n/c	militar	ACC, APP, TWR	ACC	APP, TWR	GRP, TWR, APP, ACC	s/d
c) país	US	s/d	BE	FR	AL, CH, FR	CR	CH, LI, FI	FR
d) aeropuertos/centros	DFW, BOS, DCA	n/c	n/c	s/d	EUROCONTROL Experimental Centre (EEC) en Brétigny-sur-Orge, Francia	ZAG	s/d	s/d
d) tipo de interacción	operacional	texto elicitado	operacional	operacional	Simulado	operacional	operacional	operacional
Audio								
a) duración (horas)	70	s/d	24+	150	51.4	20	20	22
b) archivos	s/d	WAV	RIFF WAV	s/d	DAT/WAV	MP3	PCM	s/d
Hablantes								
a) nivel de inglés	mayoría inglés L1	otras L1 (fr, gr, it, es)	mayoría otras L1 (C: da y fr de BE; P: al, be, da y fr de BE, fr, it, es, in de CA, UK y US).	s/d	mayoría otras L1	diversas L1	s/d	L1 fr
b) género	mixto	mixto	mayoría masculino	mixto	mixto	s/d	mixto?	s/d
c) número de hablantes	desconocido (alto)	81	desconocido (alto)	desconocido (alto)	10	s/d	s/d	s/d
Otras características								
a) cantidad de emisiones	s/d	8099	9833	s/d	10078	1967 *		
a) transcripción	Ortográfica	s/d	s/d	s/d	ortográfica	ortográfica?	fonética (Arpabet)	ortográfica
b) fronteras temporales	sí	s/d	silencios removidos	s/d	sí	s/d	s/d	
c) enmascaramiento	no	s/d	no	s/d	no	no?	s/d	sí?
d) etiquetado/anotación	s/d	s/d	hablantes e indicativos de llamada	s/d	XML-tags para otras lenguas/fases no transaccionales de la interacción	s/d	sí, XML	POS-tagging manual, etiquetado sintáctico y léxico manual
d) software de grabación/ transcripción	s/d	s/d	s/d	s/d	*TableTrans*	*Goldwave*	*WebTransc*	s/d
e) software de análisis	s/d	s/d	s/d	s/d	s/d	*WordSmith Tools 4*	s/d	s/d
f) organización interna	1 subcorpora / aeropuerto	s/d	s/d	s/d	s/d	s/d	s/d	s/d
Distribución								
a) disponibilidad	sí (paga)	sí (?, a pedido)	restringido a miembros OTAN, autorizable a otros investigadores	no	sí (gratuita)	s/d	s/d	s/d
b) formato	CD-ROM	s/d	s/d	s/d	DVD o imagen de disco ISO, en línea	s/d	s/d	s/d
c) distribuidor	*Linguistic Data Consortium*	s/d	prohibida	s/d	Graz University of Technology	s/d	Repositorio LINDAT CLARIN	s/d
Responsables	*Texas Instruments*, bajo contrato para *DARPA/ National Institute of Standards and Technology*	s/d	Grupo de tareas NATO RTO IST031-RTG013	Centre d'Études de la Navigation Aérienne' (CENA/Services de la Navigation Aérienne' (DSNA)	Graz University of Technology/Signal Processing and Speech Laboratory	s/d	Agencia de Tecnología de la República Checa, proyecto "Intelligent technologies for improving air traffic security"	French Civil Avition University (ENAC)

CLAVES: ASR: reconocimiento automático de habla // C: controlador aéreo; P: piloto // ACC: control de área/en ruta APP: aproximación; GRP: tierra; TWR: torre; be: belga; fr: francés; gr: griego; in: inglés; it: italiano; es: español; da: danés; AL; Alemania; BE: Bélgica; CA: Canadá; CH: República Checa; CR: Croacia; FI: Filipinas; FR: Francia; LI: Lituania; UK: Inglaterra; US: Estados Unidos.

NOTAS: * medido en intercambios, no en emisiones.

Tabla n.° 1. Comparación de los *corpus* de habla ATC. (Ampliado de Hofbauer et al. 2008: 13)

3. Propuesta para un *Corpus de Interacciones ATC en Argentina* (CIATCA)

3.1 Objetivos y antecedentes de la propuesta

Nuestra propuesta se diferencia de las reseñadas porque consiste en conformar un *corpus* de datos orales provenientes de interacciones reales dentro del espacio aéreo de la Argentina, tanto en inglés como en español —las dos lenguas que se registran en las frecuencias de radio en nuestro país y que determinan un espacio aéreo parcialmente bilingüe en determinados sectores—. Su finalidad consiste en el análisis de los patrones recurrentes en ambas lenguas, para generar un conocimiento más amplio sobre las diferencias existentes en el comportamiento lingüístico —fraseológico o no— de los profesionales aeronáuticos argentinos en cada una de ellas y así mejorar los procesos de instrucción comunicativa en ambas lenguas. Asimismo, el CIATCA permitirá continuar valorando la relevancia de los diversos rasgos –reglamentarios/fraseológicos o no– para la seguridad lingüística de la actividad[3].

Esta propuesta surge como necesidad tras la valoración crítica de nuestra labor previa (de-Matteis 2009), que fue realizada, en una primera etapa, con el apoyo de una base de datos diseñada originalmente en *Microsoft Access*. Durante esta instancia pudimos constatar que el análisis de los datos con esta herramienta es dificultoso y lento. La base de datos consta de 30 tablas relacionadas y 11 formularios para el ingreso de los datos correspondientes a los distintos niveles de análisis lingüístico (figura 1). Analizar cada transmisión en cada uno de ellos es una tarea que insume un tiempo considerable y la visualización de los distintos formularios para el ingreso de los datos es limitada y no permite una visión integral de los fenómenos asociados con cada transmisión. Trabajar con SQL (*Structured Query Language*) para confeccionar consultas sobre los datos almacenados, por otra parte, no resulta una tarea sencilla ni intuitiva para un usuario sin conocimientos previos y los asistentes (*wizards*) propios de *Access* no permiten explotar todo el potencial del lenguaje SQL, ya que siguen una serie de patrones definidos previamente para automatizar las consultas más habituales.

[3] Para ello se gestionarán las autorizaciones correspondientes con las autoridades y explotadores aeronáuticos para el uso enmascarado de los datos con el propósito de ajustarnos en el futuro a la Ley n.° 26899 sobre repositorios digitales y evaluar la posibilidad de hacer del *corpus* una herramienta abierta para usos académicos.

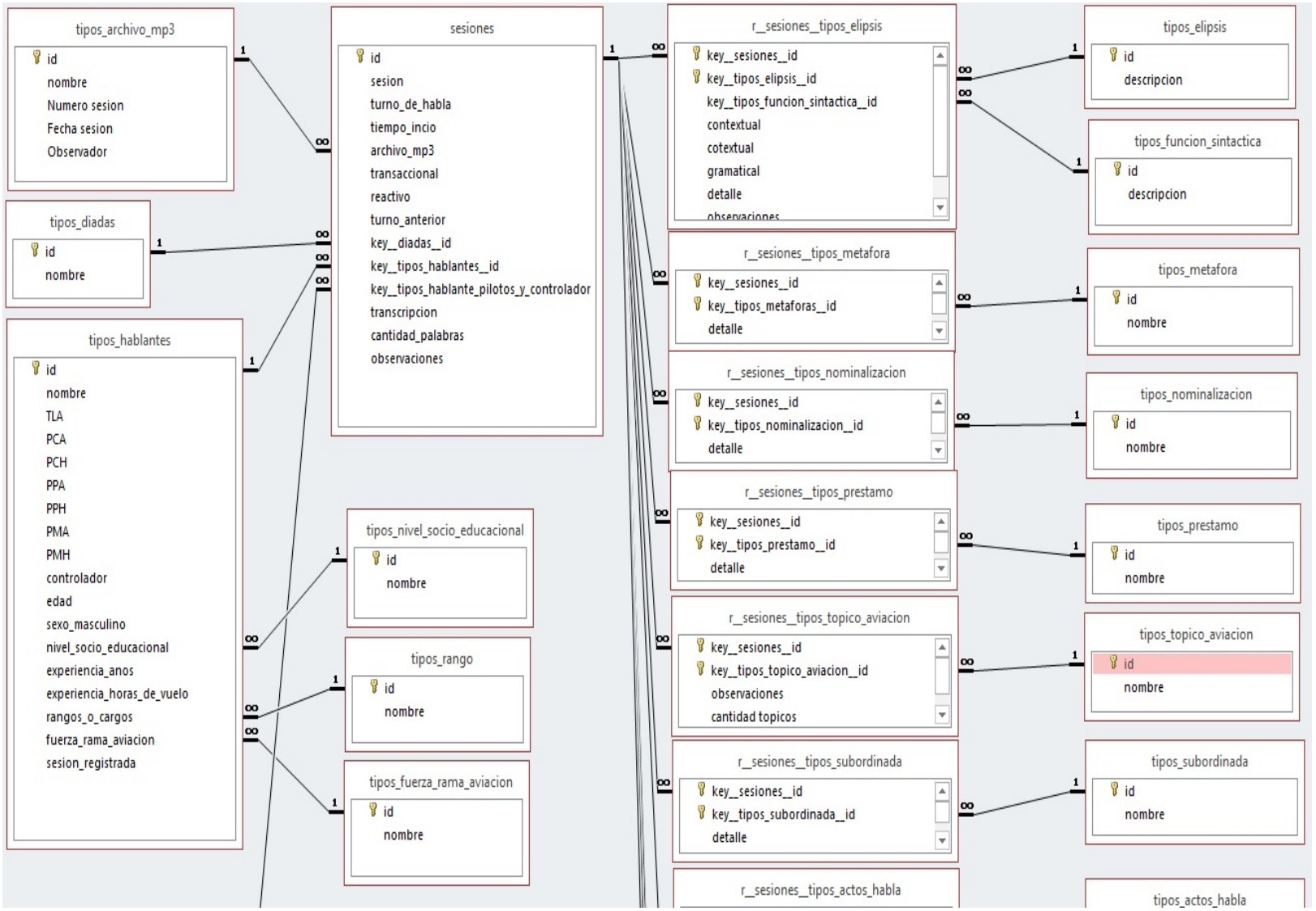

Figura n.° 1. Algunas relaciones entre tablas en la base de datos en Microsoft Access (de- Matteis, 2009).

Como alternativa superadora, intentamos trasladar este trabajo a un archivo de texto para su procesamiento con *WordSmith Tools*, conjunto de herramientas (*Concord*, *Keywords* y *WordList*) que permite trabajar con *corpus* etiquetados o no (Berber Sardinha, 2009). Para ello, desarrollamos un sistema de etiquetado que considerara los distintos aspectos relevantes para el tipo de interacción bajo estudio, desde las descripciones más generales de cada sesión de observación, a la definición de hablantes y turnos, así como de los rasgos sintácticos, léxicos, pragmáticos y discursivos más relevantes (tabla n.° 2).

Nivel	Tag	Descripción
Descriptores	<header>, </header>	Información general
	<sph>, </sph>	Esfera aeronáutica
	<oper>, </oper>	Operación
	<range>, </range>	Alcance
	<obs>, </obs>	Observaciones
Etiquetado pragmático: identificación de hablantes	<VOCR>, </VOCR>	Vocativo reglamentario
	<VOCS>, </VOCS>	Vocativo social
	<REFR>, </REFR>	Referencial reglamentario
	<REFS>, </REFS>	Referencial social
	<AUTOR>, </AUTOR>	Autorreferencial reglamentario
	<AUTOS>, </AUTOS>	Autorreferencial social
	<tuteo>, </tuteo>	Pauta de tratamiento tuteante
Etiquetado pragmático: actos de habla	<INS>, </INS>	Instrucción
	<FREC>, </FREC>	Transferencia de frecuencia
	<AUTO>, </AUTO>	Autorización explícita
	<PROH>, </PROH>	Prohibición
	<INFO>, </INFO>	Información (meteorológica, de posición, etc.)
	<COL>, </COL>	Colación (de pilotos y/o controladores
	<PREG>, </PREG>	Pregunta directa
	<SOL>, </SOL>	Solicitud
	<REC>, </REC>	Recepción
	<ID>, </ID>	Identificación en radar
	<SALUDO>, </SALUDO>	Saludos
	<AGRAD>, </AGRAD>	Agradecimientos
	<DES>, </DES>	Expresión de deseos
Etiquetado léxico	<NOM>, </NOM>	Nominalización
	<PREST>, </PREST>	Préstamo
	<SIGL>, </SIGL>	Sigla
	<QQQ>, </QQQ>	Elemento código Q
	<MET>, </MET>	Metáfora

Tabla n.° 2. Muestra de algunas de las etiquetas propuestas para el etiquetado del *corpus* P/C (de- Matteis, 2012).

El trabajo con muestras parciales tomadas de nuestro *corpus* previo, así como con muestras de interacción P/C más recientes, nos convenció de que esta alternativa resulta mucho más productiva. Sin embargo, nos encontramos con que no teníamos una herramienta que permitiera simplificar el etiquetado y que realizarlo de forma manual introducía amplias posibilidades para el error, además de insumir también demasiado tiempo sin una herramienta que asistiera en el proceso para garantizar su consistencia. Cabe aclarar que *WordSmith Tools* no provee una herramienta para simplificar el

etiquetado y, al momento de desarrollar nuestras etiquetas no conocíamos ni manejábamos el estándar provisto por la *Text Encoding Initiative* (TEI) para la transcripción de textos orales (TEI, 2015).

3.2 Necesidad de una herramienta específica

De acuerdo con esta breve reseña de nuestra labor previa, nos encontramos utilizando simultáneamente varias herramientas independientes (un procesador de textos —cualquiera de ellos— para transcribir y etiquetar manualmente, un reproductor de audio, un gestor de *corpus* lingüísticos que, a su vez, es multimodular) y sin poder aún automatizar una serie de operaciones necesarias para el análisis de nuestros datos.

La experiencia con el *software* mencionado hasta este punto nos llevó a investigar, en primer lugar, sobre la existencia de programas integrados para la transcripción y etiquetado de *corpus*[4]. Tras revisar distintos tipos de aplicaciones, encontramos proyectos en desarrollo en otras universidades, muchos de los cuales ya no cuentan con mantenimiento, como también varios sistemas propietarios. Estos últimos pueden ser del tipo *standalone* (es decir, que se ejecutan en sistemas operativos GNU/Linux o bien en *Microsoft Windows*) o presentarse como sistemas en la *web* (v. 3.3.1). Pero ninguno de ellos permite abordar de manera ágil las especificidades de la interacción que ocurre en el ámbito ATC lo que nos llevó a concluir que resulta necesaria una herramienta especializada que permita aprovechar la característica rutinización de estas interacciones para agilizar, en primera instancia, el proceso de transcripción. Nos referimos, por ejemplo, a explotar la funcionalidad de texto predictivo con un sistema basado en los elementos de la fraseología estandarizada y el léxico más habitual en el lenguaje aeronáutico no fraseológico (el componente "llano" de las variedades lingüísticas para los fines aeronáuticos). Al mismo tiempo, nos planteamos la necesidad de contar con un sistema propio especializado en la transcripción de interacciones ATC que brinde opciones adicionales no disponibles en otras herramientas ya no solo para la transcripción y etiquetado de nuestros datos sino para automatizar también otro tipo de operaciones que el procesamiento tanto cuali- como cuantitativo de la interacción ATC requiere: enmascaramiento aleatorio pero consistente de los datos que permitieran identificar vuelos, explotadores y hablantes de las distintas organizaciones; posibilidad de indicar y recuperar elementos elididos; filtrado de las interacciones objeto de estudio respecto de otras que se registran alternadas con las primeras en la misma frecuencia de radio para su análisis individual; visualización de cada hablante y organización particular con distintos colores; visualización de las secuencias interaccionales vinculadas a cada fase de vuelo; exportación de los datos en distintos formatos, entre otras.

3.3 Descripción de la propuesta

Se propone desarrollar un sistema modular y por capas (figura n.° 2) que permita integrar las herramientas para la conformación del *corpus* (Transcriptor y Etiquetador) con las dedicadas a su

[4] Aunque su uso no está extendido en la lingüística de *corpus* (Rayson, 2015), se consideraron también aplicaciones que posibilitan el análisis asistido de datos en modo cualitativo, como *Atlas.ti*.

análisis (Analizador). En dicha empresa, se aprovecharán herramientas disponibles de código abierto siempre que sea posible, introduciendo las modificaciones necesarias para satisfacer las necesidades específicas impuestas por nuestro objeto de estudio.

3.3.1 Transcriptor

Se trabajará sobre la base del *software* disponible para la transcripción de texto oral, integrando las funcionalidades ofrecidas por diversos sistemas como *Transcriber*[5], los provistos por el *Linguistic Data Consortium* (como *XTrans*[6] o el *Annotation Graph Tool Kit* [AGTK][7]), las herramientas del *UAM Corpus Tool*[8], *ELAN*[9], *EXmeralda*[10] o el *software* propietario *WebTransc* [Šmídl e Ircing, 2014] e, incluso, los programas para subtitulado adaptados a fines lingüísticos como *Subtitle Workshop*[11]. Se incluirá la posibilidad de realizar un etiquetado inicial para las circunstancias contextuales que rodean a la interacción (esto es, ruidos ambientales, interferencias en las frecuencias de radio, fragmentos ininteligibles), como así también de definir con claridad a los distintos hablantes, estableciendo un código de colores para sus turnos de habla. Asimismo, se prevé también contar con una ayuda visual con las convenciones de transcripción adoptadas para complementar la transcripción ortográfica de las interacciones indicando —solo cuando su presencia resulta significativa para el análisis de la interacción ATC— algunos de los rasgos fonético-fonológicos que pueden resultar más salientes en una determinada emisión[12].

3.3.2 Etiquetador

Dado que los objetivos de análisis de nuestra investigación no requieren *POS-tagging* —forma de etiquetado que podría automatizarse mediante algoritmos específicos pero que resultaría poco operativa ya que nuestras muestras incluyen intervenciones tanto en inglés como en español—, se definirá un esquema de etiquetas basado en el modelo TEI y otros desarrollos que nos servirán como orientación (Schmidt, 2011). A estas etiquetas, se añadirán aquellas que resultan específicas del tipo de interacción bajo estudio en los distintos niveles de análisis lingüístico. El etiquetado será, entonces, manual pero se proveerá de herramientas visuales para agilizar su proceso —al estilo de los procesadores de texto más familiares— y de métodos para garantizar su consistencia.

[5] http://xml.coverpages.org/transcriber.html [accedido el 22/8/2016].
[6] https://www.ldc.upenn.edu/language-resources/tools/xtrans [accedido el 22/8/2016].
[7] http://agtk.sourceforge.net/, [accedido el 22/8/2016].
[8] http://www.corpustool.com/index.html [accedido el 22/8/2016].
[9] https://tla.mpi.nl/tools/tla-tools/elan/ [accedido el 22/8/2016].
[10] http://www.exmaralda.org/en/downloads/ [accedido el 22/8/2016].
[11] Disponible en http://subworkshop.sourceforge.net/index.php y cuyo tutorial se puede consultar en http://wiki.linguisticteam.org/w/Subtitle_workshop_tutorial; o bien la adaptación disponible en http://www.uruworks.net/index.html (accedidos todos el 22/8/2016).
[12] Teniendo en cuenta la imposibilidad de realizar transcripciones finas en los niveles fonético-fonológico debido a la calidad de las grabaciones disponibles, una solución de compromiso que hemos adoptado en trabajos previos es la de incorporar a las convenciones de transcripción de Schegloff y Sacks (2000) —que permiten dar cuenta al menos parcial de rasgos como el volumen y el énfasis de las emisiones— la transcripción fonológica de unidades léxicas particulares —por ejemplo, en el caso de los préstamos— o de frases que resultan de interés en este nivel, incluyendo en esta transcripción rasgos fonéticos como el alargamiento vocálico.

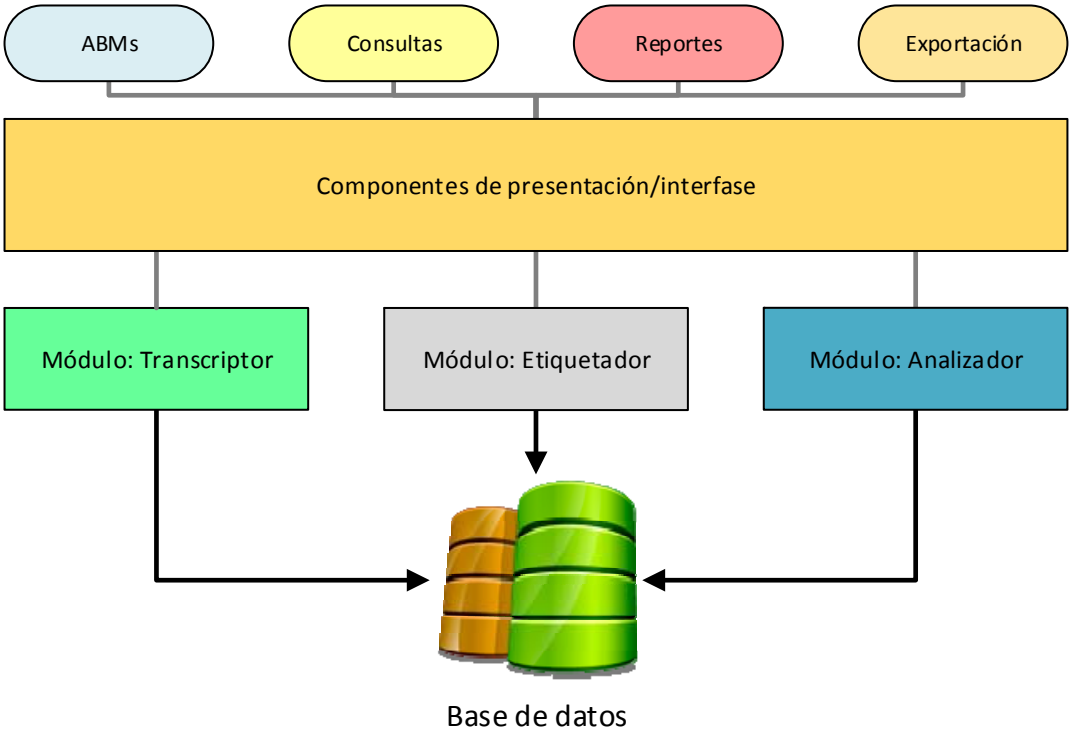

Figura n.° 2. Esquema del sistema propuesto.

3.3.3 Analizador

Los datos que componen el *corpus*, las transcripciones y sus versiones etiquetadas, serán procesados mediante una serie de programas que se desarrollarán aprovechando el conjunto de librerías NLTK para Python, herramienta ampliamente utilizada en la lingüística de *corpus*. Estos programas conformarán la capa central del sistema, a la que denominamos provisionalmente como *Analizador.py*. Los módulos y librerías que se desarrollarán permitirán realizar las tareas automáticas de análisis típicas de la lingüística de *corpus*: cuantificar *tokens* y *types*, elaborar listas de palabras con sus respectivas frecuencias, lematizar, elaborar concordancias, identificar colocaciones, etc., así como relacionar fenómenos específicos de la interacción ATC —definidos en los distintos niveles de etiquetado—, tanto entre sí como con las características de los participantes de la interacción y de las circunstancias operacionales. Por ejemplo, se realizarán operaciones como: cuantificar casos y clases de elementos elididos, tipo y frecuencia de cláusulas subordinadas, clases de elementos léxicos, niveles de personalización del discurso (a través de la consideración de las frecuencias de uso de las diversas fórmulas de tratamiento, pronombres personales y formas verbales conjugadas, etc.) y correlacionar todas estas medidas con los distintos colectivos profesionales de acuerdo con variables sociales específicas.

3.3.4 Interfaz

La interfaz hombre-máquina se diseñará de manera que la operación del sistema resulte lo más transparente e intuitiva posible, atendiendo a la posibilidad de que pueda ser empleada por investigadores con distinta experiencia en el empleo de sistemas informáticos y tratando de evitar, así, las

dificultades ya experimentadas en el trabajo previo sobre aplicaciones de bases de datos sin perder flexibilidad y poder analítico. Se propone presentar al usuario un sistema que conste de diferentes módulos accesibles de manera independiente, posibilitando dividir las tareas de los investigadores en base a los tres componentes definidos hasta acá: Transcriptor, Etiquetador y Analizador, además de agregar un módulo específico para obtener reportes a partir del procesamiento del *corpus* de trabajo.

4. Ventajas de la propuesta y proyecciones futuras

Además de las posibilidades de análisis inherentes que ofrecerá un sistema especialmente orientado hacia la investigación de la interacción ATC, tanto en escenarios mono como bilingües, las ventajas de contar con un sistema que integre las distintas herramientas con las que, hasta la fecha, venimos trabajando de manera independiente, permitirá secuenciar de manera más ordenada las tareas de elaboración de los datos —transcripción, etiquetación progresiva en los distintos niveles de análisis requeridos, análisis cuantitativo, análisis cualitativo—, como así también realizar un seguimiento del grado de avance en la preparación y procesamiento de los materiales. Esto establece una diferencia con respecto a los *corpus* ATC reseñados, puesto que las herramientas empleadas en la mayoría de ellos tampoco están integradas.

La labor desarrollada hasta la fecha en la constitución de un *corpus* y la definición precisa de las necesidades nos permite proyectar una labor conjunta entre el Departamento de Humanidades y el Departamento de Ciencias e Ingeniería de la Computación. Esto permitirá incentivar la investigación interdisciplinaria tanto al nivel de docentes-investigadores como en proyectos parciales que podrían ser implementados por estudiantes de grado interesados en distintos aspectos vinculados al procesamiento del lenguaje natural, posiblemente involucrando a estudiantes de las carreras de Ingeniería en Sistemas de Información y Licenciatura en Ciencias de la Computación, con formación para el desarrollo de *software* y capacidades para realizar trabajos relacionados con otras áreas de conocimiento.

Por otra parte, como lo evidencian los antecedentes en materia de *corpus* ATC, el diseño modular y de capas propuesto constituirá una propuesta innovadora en el campo específico del procesamiento de los datos ATC en el mundo hispanoparlante, que podrá eventualmente ponerse a disposición de otros investigadores como herramientas utilizables en línea. En etapas futuras de desarrollo, por otra parte, podrán ponerse a disposición de otros científicos los programas compilados en Python correspondientes a la capa del Analizador, de manera que puedan hacer uso de las mismas sobre los datos exportados (previamente cargados en el sistema y enmascarados), o bien sobre datos generados propios, que respeten los formatos de entrada de dichos programas compilados.

5. Observación final

Por último, nos interesa realizar una reflexión final de carácter (auto)crítico a partir de la experiencia analizada y del proyecto formulado. Como hemos intentado mostrar, el camino recorrido para dar con la mejor herramienta —o conjunto de herramientas— para conformar un *corpus* ATC, si bien ha resultado enriquecedor ha estado marcado también por avances y retrocesos que, a la luz de las tecnologías disponibles actualmente, resultan innecesarios. En este sentido, y en el marco de las reflexiones sobre las prácticas de investigación en la universidad que fomentan estas jornadas, consideramos que la incorporación de las herramientas digitales en la formación de grado y de posgrado de los investigadores en

ciencias humanas resulta necesaria para minimizar el tiempo de exploración y experimentación individual, que ralentiza los procesos de examen de los datos y puede condicionar la productividad y fertilidad de los proyectos de investigación en los estudios lingüísticos. En este sentido, en el marco de unas jornadas que apuntan a reflexionar sobre nuestras prácticas de investigación y considerando el importante lugar que el procesamiento del lenguaje natural tiene también en las ciencias de la computación, sería conveniente articular, a nivel institucional, las oportunidades para el intercambio de saberes y la colaboración fructífera entre ambas ciencias.

Bibliografía

Baker, P. (2010). *Sociolinguistics and corpus linguistics*, Edinburgh, Edinburgh University Press.

Baker, P.; Hardie, A. y T. McEnery (2006) A glossary of corpus linguistics, Edinburgh, Edinburgh University Press.

Barras, C.; Geoffrois, E.; Wu, Z. y Liberman, M. (2001). "Transcriber: development and use of a tool for assisting speech corpus production", *Speech Communication*, 33, pp. 5- 22.

Berber Sardinha, T. (2009). *Pesquisa em lingüística de corpus com WordSmith Tools*, Campinas, Mercado de Letras.

Bird, S.; Klein, E. y Loper, E. (2009). *Natural language processing with Python*, Sebastopol, O'Reilly.

de- Matteis, L. M. A. (2012). *Procesamiento de corpus oral con WordSmith Tools*. Documento de trabajo inédito.

de Matteis, L. M. A.; Martino, A. y N. Bravo (2019). "Inglés aeronáutico: definición y reseña de su enseñanza en Argentina", Actas VI Jornadas de Investigación en Humanidades, Departamento en Humanidades, Departamento de Humanidades, Bahía Blanca, EdiUNS, vol. 13, pp. 815-822.

Godfrey, J. (1994). *Air Traffic Control Complete, Linguistic Data Consortium*, Philadelphia. Disponible en: https://catalog.ldc.upenn.edu/LDC94S14A. Consultado el: 9/11/2015.

Graglia, L.; Favennec, B. y Arnoux, C. (2005) "VOCALISE: Assessing the impact of Data Link Technology on the R/T channel", *The 24th Digital Avionics Systems Conference*, Vol. 1, pp. 5.C.2-51-13. Disponible en: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?reload=true&arnumber=1563381. Consultado el: 11/11/2015.

Hofbauer, K.; Petrik, S. y Hering, H. (2008). "The ATCOSIM Corpus of Non-Prompted Clean Air Traffic Control Speech", en: Calzolari, N.; Choukri, K.; Maegaard, B.; Mariani, J.; Odijk, J.; Piperidis, S. y Tapias, D. (Eds.). *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (LREC'08), Marrakech, European Language Resources Association (ELRA), pp. 2147-2152.

Horn, F. (s/f). *UAM Corpus Tool – How to*. Disponible en: http://www.linguisticweb.org/lib/exe/fetch.php?media=linguisticsweb:tutorials:linguistics_tutorials:manual_annotation:uam_corpus_tool.pdf. Consultado el: 9/11/2015.

Maeda, K. y Strassel, S. (2004). "Annotation tools for large-scale corpus development: using AGTK at the Linguistic Data Consortium", *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, pp. 2077-2080. Disponible en: http://papers.ldc.upenn.edu/LREC2004/AGTK.pdf. Consultado el: 9/11/2015.

McEnery, T. y Hardie, A. (2012). *Corpus linguistics*, Cambridge, Cambridge University Press.

O'Donnell, M. (2008). "The UAM CorpusTool: software for corpus annotation and exploration", *Actas del XXVI Congreso de AESLA*, Almería.
Disponible en: https://www.uam.es/proyectosinv/woslac/DOCUMENTS/Presentations% 20and %20articles/ODonnellAESLA08.pdf. Consultado el: 9/11/2015.

O'Keeffe, A. y McCarthy, M. (Eds.) (2010). *Handbook of* Corpus *Linguistics*, New York, Routledge.

Pavlinović, M.; Boras, D. y Francetić, I. (2013a). "First steps in designing Air Traffic Control communication language technology system - Compiling spoken corpus of radiotelephony communication", *International Journal of Computers and Communications*, vol. 3, n.º 7, pp. 73- 80.

Pavlinović, M.; Boras, D. y Juričić, B. (2013b). "Spoken Corpus of radiotelephony phraseology", en: Boras, D.; Mikelic Predarovic, N.; Moya, F.; Roushdy, M. y Salem, A.-B. M. (Eds.). *Recent Advances in Information Science, Proceedings of the 7th European Computing Conference* (ECC '13), Dubrovnik, WSEAS Press, pp. 136-141.

Pigeon, S.; Shen, W. y van Leeuwen, D. (2007). "Design and characterization of the non-native military air traffic communications database (nnMATC)", *International Conference on Spoken Language Processing* (INTERSPEECH), Antwerp, Bélgica.

Rayson, P. (2015). "Computational tools and methods for *corpus* compilation and analysis", en: Biber, D. y Reppen, R. (Eds.). *The Cambridge Handbook of English* Corpus *Linguistics*, Cambridge, Cambridge University Press, pp. 32-49.

Schmidt, T. (2011). "A TEI-based approach to standardizing spoken language transcription", *Journal of the Text Encoding Initiative*, 1, pp. 2-22.

Segura, J. C.; Ehrette, T.; Potamianos, A.; Fohr, D.; Illina, I.; Breton, P-A.; Clot, V.; Gemello, R.; Matassoni, M. y Maragos, P. (2007). "The HIWIRE database, a noisy and non-native English speech *corpus* for cockpit communication".
Disponible en: http://cvsp.cs.ntua.gr/projects/pub/HIWIRE/HiwirePublications/HIWIRE_db_description_paper.pdf. Consultado el: 9/11/2015.

Šmídl, L. y Ircing, P. (2014). "Air Traffic Control Communication (ATCC) Speech *Corpus*", paper presentado en CLARIN Annual Conference, Soesterberg, Países Bajos, 23-25 de octubre.
Disponible en: https://www.clarin.eu/sites/default/files/cac2014_submission_1_2.pdf. Consultado el: 9/11/2015.

Stulic-Etchevers, A. y Rouissi, S. (2009). "Pensando un *corpus* en modo colaborativo: hacia el prototipo del *corpus* judeoespañol digital", en: Enrique-Arias, A. (Ed.). *Diacronía de las lenguas iberorrománicas. Nuevas aportaciones desde la lingüística de* corpus, Madrid, Iberoamericana-Vervuert, pp. 117-134.

Text Encoding Initiative (2015). *TEI P5: Guidelines for Electronic Text encoding and interchange*.
Disponible en: http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf. Consultado el: 9/11/2015.

Villayandre Llamazares, M. (2010). *Aproximación a la lingüística computacional*, León, Universidad de León.

Viudas Camarasa, A. (1990). "Inteligencia artificial en filología", *Anuario de Estudios Filológicos*, 13, pp. 403-409.