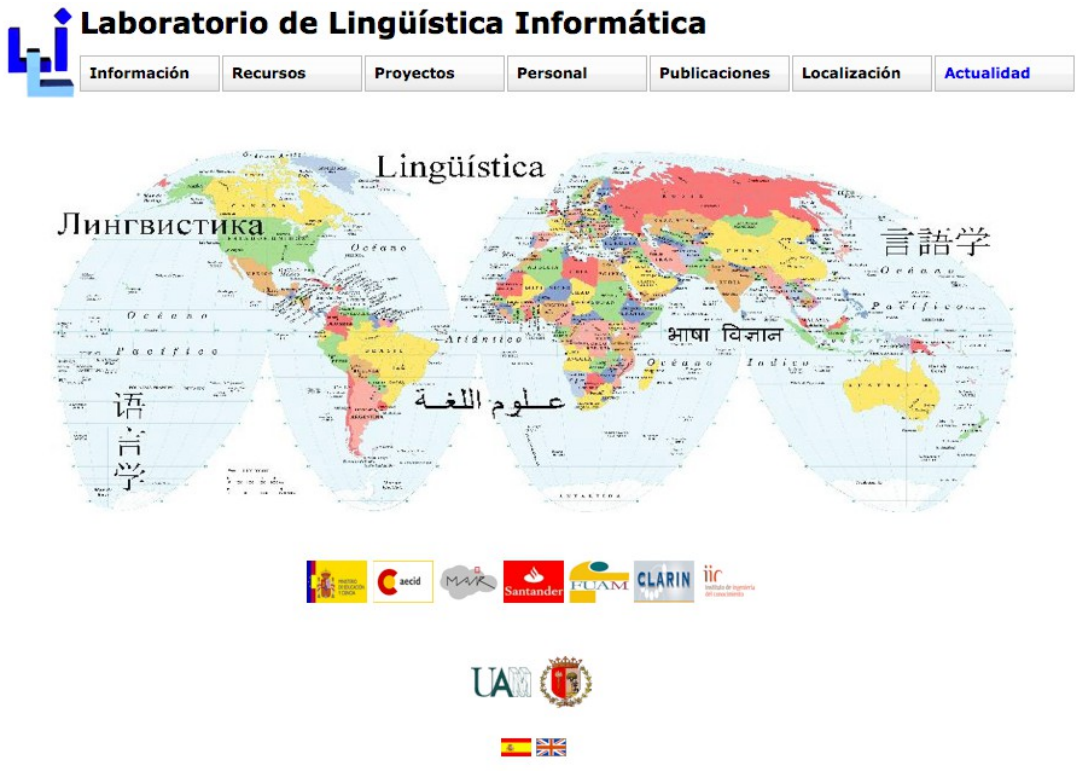
El Laboratorio de Lingüística Informática de la Universidad Autónoma de Madrid

Antonio Moreno Sandoval

El Laboratorio de Lingüística Informática (LLI-UAM, http://www.lllf.uam.es) es un grupo de investigación reconocido por la UAM que tiene sus orígenes a mediados de los años 80. Fundado por el Prof. Francisco Marcos Marín con un grupo de estudiantes de doctorado, a partir de su experiencia en el Centro de Investigación de IBM en España. La fundación como tal del LLI se produjo con la participación del grupo en el proyecto europeo de traducción automática EUROTRA (diciembre 1987 a diciembre de 1992). Desde entonces, el grupo cuenta con espacio permanente en la Facultad de Filosofía y Letras de la UAM (http://www.lllf.uam.es/ESP/images/mapalab.jpg), un servidor web propio y un técnico informático como administrador de sistemas.



La primera etapa del Laboratorio, bajo la dirección del Prof. Marcos Marín, se repartió entre dos líneas de investigación: las humanidades digitales y la lingüística computacional basada en corpus. Gracias al patrocinio de la Sociedad Estatal del Quinto Centenario, entre 1990 y 1992 se compilaron recursos digitales pioneros en

España como ADMYTE (Archivo Digital de Manuscritos y Textos Electrónicos) o CORLEC (Corpus Oral de Referencia de la Lengua Española Contemporánea). En la línea más computacional, dentro de la década de los noventa, se realizó la parte española del proyecto europeo CRATER (Corpus Resources And Terminology ExtRaction, 1994-95), que produjo un corpus paralelo trilingüe de un millón de palabras, anotado morfosintácticamente y alineado a nivel de oración. También destacan los proyectos en corrección gramatical automática (GramCheck y CONT-TEXT), dirigidos por Fernando Sánchez y Flora Ramírez. Entre 1997 y 2000 se compiló el primer corpus anotado sintácticamente del español, el *UAM Spanish Treebank*, financiado por la New York University y dirigido por Antonio Moreno Sandoval. Paralelamente, se realizaron otros proyectos menores con distinta financiación (CICYT, Anaya, Acciones Integradas Hispano-Alemanas) centrados en diccionarios electrónicos. Desde 1996, Fernando Sánchez empezó a colaborar con la RAE en la creación del corpus CREA y en 2001 se incorporó definitivamente como director del Departamento de Tecnologías Lingüísticas de la institución académica. Un año antes, en 2000, Flora Ramírez dejó el LLI para unirse al grupo de lingüística computacional en español de Microsoft, en Redmond, Estados Unidos. En esos mismos años, el Prof. Marcos Marín ocupó el puesto de Director Académico del Instituto Cervantes y posteriormente una cátedra en Roma-La Sapienza, lo que en la práctica supuso su ausencia en la participación en los proyectos del grupo hasta su jubilación en 2006. La vinculación del fundador con el LLI se ha mantenido todos estos años.

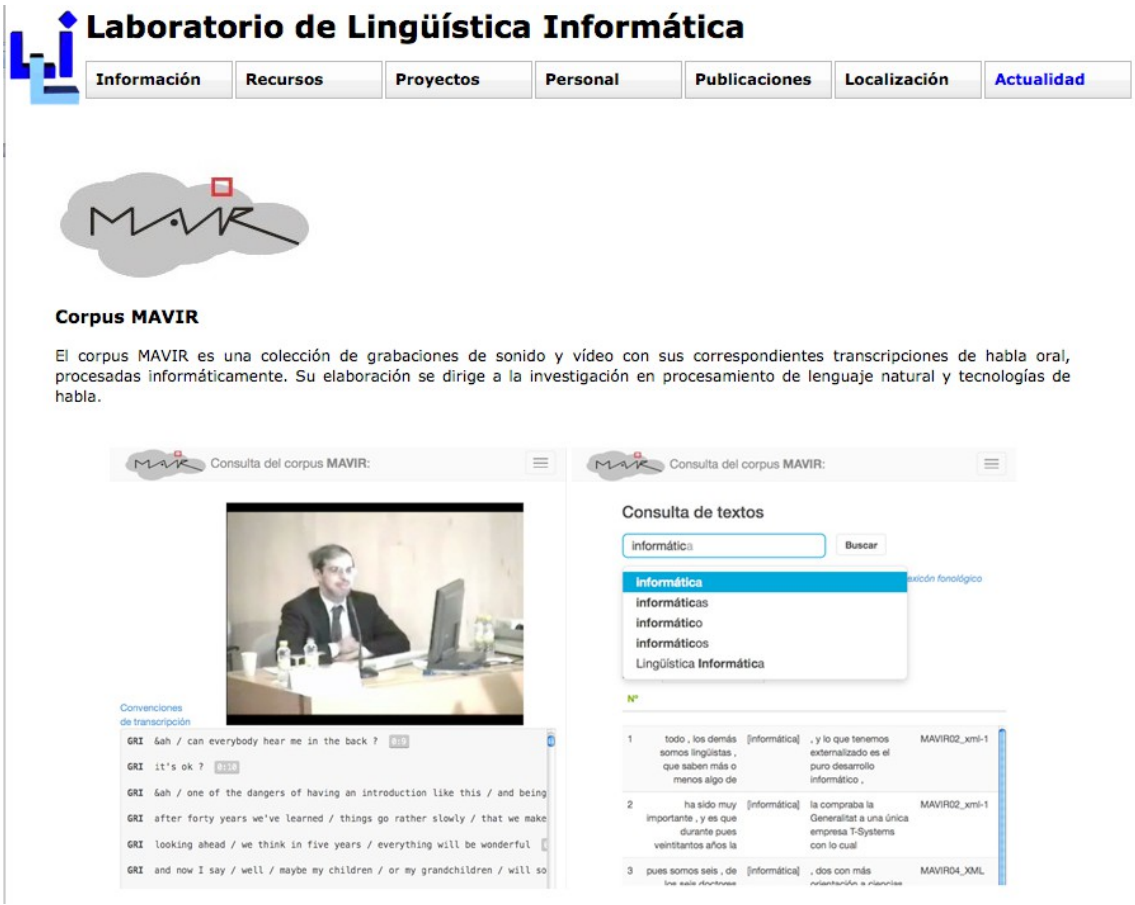
La segunda etapa del LLI-UAM comenzó con el proyecto europeo C-ORAL-ROM (2001-2004) ya bajo la dirección de Antonio Moreno Sandoval. Este corpus oral supuso un punto de inflexión en los recursos elaborados por el LLI y ha marcado el modelo de los otros corpus elaborados posteriormente (Moreno Sandoval 2002 describe la evolución entre CORLEC y C-ORAL-ROM). El grupo de becarios que se formó en el proyecto europeo luego continuaron con tesis doctorales que expandieron la anotación del corpus original (en concreto, análisis semántico eventivo, Manuel Alcántara, y pragmático discursivo, Ana González) o crearon nuevos corpus: CHIEDE (corpus de habla infantil, Marta Garrote), CORELE (corpus oral de aprendices de español, Leonardo Campillos), CORAF (corpus oral de aprendices de francés, Ana Valverde). Al tiempo que se desarrollaban estos recursos en español, el Laboratorio se internacionalizó con la incorporación de becarios extranjeros con ayudas de la AECID para realizar sus tesis en España: Doaa Samy, procedente de la Univ. de El Cairo, inauguró en 2005 las tesis en otras lenguas, en concreto el árabe, que se convertiría desde entonces en una de las lenguas de trabajo del LLI. Alicia González desarrollaría posteriormente un analizador morfológico del verbo en árabe (en 2013) siguiendo la tradición de los analizadores morfológicos iniciada por Moreno y Sánchez en los 90. Las lenguas china y japonesa se incorporaron después al catálogo de corpus orales (C-ORAL-CHINA y C-ORAL-JAPON, respectivamente), que dieron lugar a las tesis de Yang Dong (2012) y Emi Takamori (2014).

El LLI-UAM ha desarrollado una metodología de elaboración de corpus de habla espontánea como un proceso sistematizado y se ofrece como un servicio de asistencia técnica entre el cliente y el LLI, gestionado a través de la Fundación UAM. El trabajo incluye todas las etapas del desarrollo desde el diseño del corpus, la

captura de los datos y el posterior análisis, anotación y enriquecimiento de la colección.

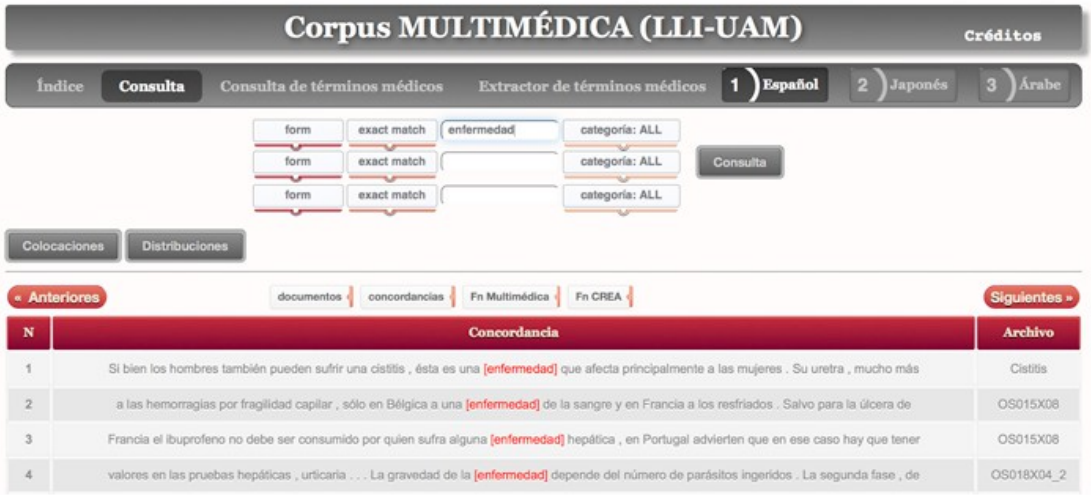
- Diseño preliminar teniendo en cuenta las características socio-lingüísticas (edad, sexo, datos demográficos, origen lingüístico, educación, etc.) y el contexto comunicativo. Esta información puede modificarse en función de los objetivos del estudio y el diseño puede adaptarse a las variables a considerar.
- Recolección de datos (grabaciones, capturas de video, edición).
- Transcripción ortográfica (indicando tanto la variante normativa como la enunciación real).
- Anotación prosódica, marcas de pausa, alargamientos vocálicos, solapamientos, interrupciones, entonación, etc.
- Alineamiento de unidades de texto-sonido en enunciados.
- Anotación morfológica semi-automática (información morfológica y lemas), con revisión manual por especialistas.
- Anotación fonológica automática.

Un ejemplo de esta metodología se puede comprobar con el corpus MAVIR (http://www.lllf.uam.es/ESP/CorpusMavir.html), encargado por el consorcio madrileño del mismo nombre y que recoge una colección de grabaciones de sonido y vídeo de conferencias sobre temas de tecnología informática, en español e inglés, con sus correspondientes transcripciones.



En la misma línea de elaboración de recursos para tecnologías del habla, hemos recogido un corpus de grabaciones de móviles para la empresa española Sigma Technologies y ofrecemos en línea una pequeña base de datos acústica con preguntas en cuatro lenguas (español, árabe, japonés y thai).

Desde 2011 hemos vuelto a los corpus escritos: el proyecto MultiMedica (financiado por el MINECO) nos ha permitido compilar un corpus de textos médicos en español, japonés y árabe y desarrollar una herramienta de consulta en línea, que ofrece además un extractor automático de términos médicos en las tres lenguas. Hasta la fecha, es nuestro recurso más elaborado pues combina la compilación de los corpus, el procesamiento morfológico, la indexación para la consulta, el desarrollo de los repertorios terminológicos y, por último, la creación de reglas para extracción de candidatos a término. Todo ello en tres lenguas muy distanciadas genética y tipológicamente, con sistemas de escritura muy dispares. El resultado se puede consultar en http://www.lllf.uam.es/ESP/Multimed.html.



La investigación del LLI en los próximos años continuará esta línea de trabajo en corpus textuales especializados, sin abandonar las líneas en corpus orales y en lenguas asiáticas. En concreto, Carlos Herrero (becario FPI) está preparando su tesis sobre anotación de la negación y la modalidad en los corpus orales de español y japonés. Yuanyi Liu (becaria del Gobierno chino) realiza una tesis contrastiva entre español y chino dentro del contexto de la traducción.

La lista completa de proyectos realizados durante estos 25 años de existencia del LLI-UAM se puede consultar aquí. La siguiente tabla resume los recursos lingüísticos desarrollados disponibles.

RECURSO	TIPO	USO	CARACTERÍSTICAS
CORLEC **Corpus Oral de Referencia de la**	Corpus	Libre	Base de datos textual (corpus de lengua oral): 1.100.000 de palabras transliteradas en soporte informático.

Lengua Española Contemporánea			
Corpus de Referencia de la Lengua Española en la Argentina	Corpus	Libre	Base de datos textual (corpus de lengua escrita): más de 2.000.000 de palabras
Corpus de Referencia de la Lengua Española en Chile	Corpus	Libre	Base de datos textual (corpus de lengua escrita): 2.000.000 de palabras
Spanish Treebank Corpus	Corpus	Libre	1.500 oraciones extraídas de periódicos y anotadas sintácticamente
C-ORAL-ROM	Corpus	Restringido	Corpus oral multilingüe español-francés-portugués-italiano con 300.000 palabras en cada lengua
CHIEDE **Corpus de Habla Infantil Espontánea del Español**	Corpus	Libre	Corpus oral de lenguaje infantil con alrededor de 60.000 palabras
Corpus Oral de Español como Lengua Extranjera	Corpus	Libre	Corpus oral de interlengua de estudiantes de español con más de 50.000 palabras.
Corpus Oral de Aprendientes de Francés	Corpus	Libre	Corpus oral de interlengua de aprendientes de francés con más de 61.000 palabras.
GRAMPAL	Programa	Restringido	Etiquetador morfosintáctico.
Corpus Árabe-Español	Corpus	Libre	Corpus paralelo árabe-español con 1179 oraciones
Diccionario Español-Francés	Diccionario	Libre	Diccionario de dificultades de uso de las preposiciones en el idioma francés
JAPONÉS	Corpus y diccionario	Restringido	Corpus oral del japonés de unas 50.000 palabras y diccionario de las 800 palabras básicas del japonés con sonido.
Corpus MAVIR	Corpus	Restringido	Corpus oral en el que se recopilan las conferencias de las Jornadas MAVIR.
Base de datos acústica de preguntas	Base de datos	Restringido	Colección de preguntas orales recopilada a partir de la participación en el CLEF
Analizador morfológico de árabe	Programa	Libre	Demo

Colaboraciones con otros grupos de investigación

Desde su fundación, el LLI-UAM ha mantenido una estrecha vinculación con diferentes grupos nacionales e internacionales. El Centro de Investigación de IBM en Madrid y los equipos europeos del proyecto EUROTRA marcaron la primera línea de internacionalización: Marcos Marín y su relación con el centro alemán de IBM en Heidelberg; Moreno Sandoval con su participación en un proyecto con IBM Suecia; Sánchez León y Ramírez Bustamante con el grupo alemán de EUROTRA en Saarbruecken. La estancia postdoctoral de Moreno en el grupo dirigido por el prof. Ralph Grishman en la NYU durante los años 1991 y 1992 ha supuesto sin duda una productiva línea de colaboración (por ejemplo, el UAM Treebank), que ha continuado hasta el presente con estancias cortas de Grishman y Sekine en el LLI, financiadas por el consorcio MAVIR.

A mediados de los 90, los catedráticos Francisco Marcos y Reinhold Werner (Augsburgo, Alemania) promovieron la colaboración entre investigadores de ambos equipos, en temas relacionados con diccionarios electrónicos. Claudio Chuchuy visitó Madrid en varias ocasiones y Antonio Moreno pasó un verano en la universidad bávara, donde se familiarizó con la lexicografía aplicada en un centro de referencia internacional en elaboración de diccionarios. Desgraciadamente, esa colaboración no se ha mantenido en el tiempo.

El proyecto C-ORAL-ROM supuso otra nueva oportunidad de establecer colaboraciones con grupos europeos, en este caso, con las Universidades de Florencia, Aix-en-Provence y Lisboa. Durante la primera década del nuevo milenio los intercambios y estancias de investigadores de Madrid y Florencia fueron fluidos, lo que redundó en una mejor formación de los investigadores más jóvenes. Los intercambios continúan, ahora a través de programa Erasmus+.

Ya comentamos que la segunda etapa del LLI se ha beneficiado de la visita de investigadores extranjeros que han venido a realizar sus tesis o sus estancias de investigación al Laboratorio. Lugar destacado ocupa Doaa Samy, iniciadora de las investigaciones con el árabe estándar moderno. La Dra. Samy ha pasado largas temporadas en Madrid desde 2001. En 2009, Moreno y Samy organizaron conjuntamente el primer *Encuentro Hispano-Egipcio sobre procesamiento automático y recursos lingüísticos en español y árabe*, en la Universidad de El Cairo, donde reunieron investigadores de distintas universidades de España y Egipto. Como fruto de este encuentro, se logró firmar un convenio de cooperación entre la UAM y El Cairo para el intercambio de investigadores.

Las relaciones con las universidades japonesas comenzaron en 2004, cuando un equipo de la Univ. de Estudios Extranjeros de Tokio (TUFS en el acrónimo en inglés), dirigido por el Prof. Toshihiro Takagaki visitó el Laboratorio. Posteriormente, los investigadores principales de todos los grupos de C-ORAL-ROM fuimos invitados a una conferencia organizada por TUFS en 2005. En 2009-2010, gracias a un proyecto financiado por el Banco Santander y la UAM, pudimos

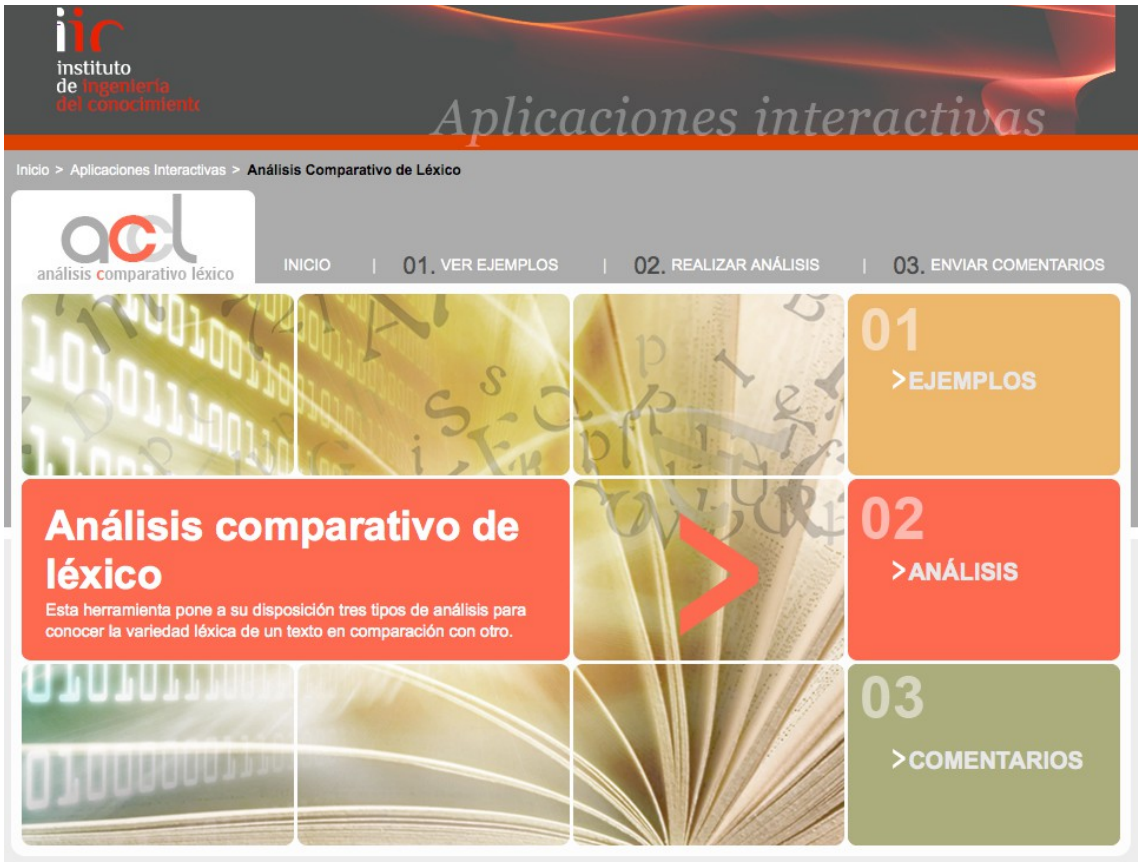
visitar Tokio para recoger las grabaciones que componen el corpus C-ORAL-JAPON. En 2013-14 hemos tenido otro proyecto similar con la Universidad de Tokio y el profesor Hiroto Ueda. En este caso, el tema del proyecto es trabajar en análisis lingüísticos sobre el español basado en recursos en formato electrónico desarrollado por ambos equipos.

La relación con las universidades chinas es más reciente. A través de la Prof. Taciana Fisac, que dirige el Centro de Estudios de Asia Oriental en la UAM, entró en contacto con nosotros Yang Dong, de Beijing International Studies University (BISU), para realizar una tesis sobre un corpus oral de chino mandarín para estudiantes españoles. La tesis, defendida en 2012, dio lugar proyecto conjunto entre UAM y BISU para elaborar materiales didácticos para la enseñanza del chino a partir del corpus. El libro está en impresión y aparecerá en 2015. Otra nueva relación, esta vez con Beijing Foreign Studies University (BFSU), se han establecido a partir de un seminario impartido por el prof. Moreno en junio de 2013 en Pekín. Desde octubre de 2014, nos visita Yuanyi Liu, profesora ayudante de BFSU, para una estancia predoctoral de tres años.

Entre los grupos españoles, debemos mencionar, por orden de antigüedad, la relación con la Dra. Núria Bel, de la UPF, con la que mantenemos relación discontinúa desde los tiempos de EUROTRA. El último contacto ha sido la infraestructura CLARIN. Más estable ha sido la colaboración con el equipo dirigido por José Carlos González, de la UPM y de la empresa tecnológica Daedalus. Un relación que comenzó en 1993 y ha continuado en numerosos proyectos conjuntos. De similar antigüedad es la colaboración con el grupo de Bases de Datos Avanzadas (LaBDA) de la UC3M, dirigido por Paloma Martínez. La Dra. Martínez en los 90 trabajó en el proyecto GramCheck con F. Sánchez y F. Ramírez. Desde 2004, el LLI y el LaBDA han participado en tres proyectos nacionales coordinados (incluyendo el proyecto MultiMedica) y dos proyectos de la Comunidad de Madrid (el consorcio MAVIR).

El LLI mantiene una fluida colaboración con diferentes investigadores y profesores de los Departamentos de Ingeniería Informática e Ingeniería de Telecomunicación en el campus de Cantoblanco. Entre otros, hemos realizado investigaciones conjuntas con Enrique Alfonseca (en Google desde 2007), Doroteo Torre Toledado, Daniel Tapias, Pablo Castells o Jordi Porta.

Desde diciembre de 2009, el LLI colabora con el Instituto de Ingeniería del Conocimiento, institución privada de I+D+i sin ánimo de lucro, sita en el campus de la UAM. Algunos de los investigadores del IIC coincidieron con Moreno en el Centro de Investigación de IBM. En la actualidad, Alicia González y Antonio Moreno investigan en el análisis de opinión y contenido en las redes sociales, así como en temas de tratamiento estadístico de textos. Un fruto interesante de esta colaboración es la aplicación gratuita Análisis Comparativo de Léxico (http://innova.iic.uam.es/acl/) que permite comparar dos textos cualquiera y extraer las palabras distintivas, su frecuencia de aparición y la riqueza léxica.



Dejo para el último lugar la colaboración más importante y fructífera de los últimos 15 años: la participación del Dr. José María Guirao, de la Universidad de Granada, como miembro del equipo en su calidad de programador senior. Su primera colaboración fue la reimplementación del analizador GRAMPAL, para convertirlo en una herramienta en línea (http://cartago.lllf.uam.es/grampal/grampal.cgi). Desde 2002 ha diseñado la estructura de los distintos interfaces de consulta a los corpus, así como la supervisión de la mayoría de los programas que se han escrito en el LLI.

Publicaciones y recursos electrónicos destacados

Filología digital

1. Marcos Marín, F. (1987): *Libro de Alexandre*. Madrid, Alianza Universidad.
[Es la primera edición unificada de una obra medieval española preparada con la ayuda de un programa informático (UNITE).]

2. Marcos Marín, F. (1994): *Informática y Humanidades*. Madrid, Gredos.
[Presenta el estado de cuestión del uso de ordenadores para estudios filológicos y lingüísticos desde mediados de los ochenta a mediados de los noventa. Es una obra que tiene interés historiográfico.]

3. Marcos Marín, F., Faulhaber, Ch., Gómez Moreno, Á. y Cortijo Ocaña, A. *ADMYTE: Archivo Digital de Manuscritos y Textos Medievales*. Micronet. (versión en línea: http://www.admyte.com/presentacion.htm).
   [Contiene las transcripciones de 290 obras redactadas en español, o en cualquiera de sus dialectos, a lo largo de la Edad Media, superando las 54.000 páginas. Es un corpus imprescindible para los estudios de historia del español, pues junto a las obras maestras como el Cantar de mio Cid o la Tragicomedia de Calisto y Melibea, se pueden consultar un catálogo sorprendente de enciclopedias, diccionarios, gramáticas, novelas de caballerías, crónicas, biografías y traducciones de clásicos grecolatinos, árabes y hebreos.]

Lingüística computacional

1. Moreno Sandoval, A. (1993): *Un modelo computacional basado en la unificación para el análisis y generación de la morfología del español*. Servicio de Publicaciones de la UAM.
   [Disponible a través de https://repositorio.uam.es/xmlui/handle/10486/12294.]

2. Moreno Sandoval, A. (1998): *Lingüística computacional: introducción a los modelos simbólicos, estadísticos y biológicos*. Madrid, Síntesis.
   [Uno de los primeros manuales sobre el tema, escrito en español. Presenta el estado de la cuestión en los 90.]

3. Moreno Sandoval, A. (2001): *Gramáticas de unificación y rasgos*. Madrid, Antonio Machado Libros.

4. Moreno Sandoval, A. y Guirao, J.M. (2006): "Morpho-syntactic Tagging of the Spanish C-ORAL-ROM Corpus: Methodology, Tools and Evaluation". En *Spoken Language Corpus and Linguistic Informatics*, John Benjamins.

5. Moreno Sandoval, A. y Guirao, J.M. *GRAMPAL: analizador morfosintáctico del español*. Versión en línea: http://cartago.lllf.uam.es/grampal/grampal.cgi.
   [Es uno de los primeros analizadores de español, con más de 20 años de desarrollo. Tiene versiones adaptadas a la lengua oral y a la escrita.]

6. González Martínez, A. (2013): *JABALIN: A computational model of Modern Standard Arabic verbal morphology based on generation*. Tesis doctoral.

https://repositorio.uam.es/bitstream/handle/10486/660335/gonzalez_martinez_alicia.pdf?sequence=1. El acceso a la aplicación es: http://elvira.lllf.uam.es/jabalin/

Lingüística de corpus

1. Moreno Sandoval, A., López, S., Sánchez, F. y Grishman, R. (2003): "Developing a syntactic annotation schema and tolos for a Spanish Treebank". En Abeillé (ed.) *Treebanks: building and using a parsed corpora*. Dordrecht, Kluwer.
   [Referencia bibliográfica donde se explica cómo se compiló y anotó el UAM Spanish Treebank.]

2. Moreno Sandoval, A, G. de la Madrid, M. Alcántara, A. González, J.M. Guirao, y R. de la Torre (2005). "The Spanish corpus". En Cresti y Moneglia (eds.) *C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages*.
   [Este capítulo describe el corpus español de C-ORAL-ROM y en especial los criterios de transcripción y anotación morfosintáctica.]

3. Campillos Llanos, L., Gozalo, P., Guirao, J.M. y Moreno Sandoval, A. (2010): *Español oral en contexto. Vol. 1. Textos de español oral. Material de ELE basado en corpus*. Madrid, Servicio de Publicaciones UAM.
   [Este libro contiene una selección de 200 fragmentos del corpus C-ORAL-ROM especialmente escogidos para realizar ejercicios de comprensión oral. En este enlace se pueden consultar algunos documentos de muestra: http://www.lllf.uam.es/ESP/CORALROM_ELE/Coralrom_ELE.html]

4. Campillos Llanos, L. "A Spanish leaner oral corpus for computer-aided error análisis". *Corpora,* 9(2): 207–238.
   [Aplicación en línea para consulta del corpus en http://cartago.lllf.uam.es/corele/home_es.html]

5. Moreno Sandoval, A., y L. Campillos Llanos (2012) "MAVIR: a corpus of spontaneous formal speech in Spanish and English".. En Torre Toledano, D., A. Ortega, A. Teixeira, J. González Rodríguez, L. Hernández Gómez, R. San Segundo, y D. Ramos Castro (eds.) *Actas de IberSPEECH 2012,* Madrid, UAM.

6. Moreno Sandoval, A., L. Campillos Llanos, C. Herrero Zorita, J. M. Guirao Miras, A. González Martínez, D. Samy y E. Takamori (2014) "An online tool for enhancing NLP of a biomedical corpus". *6º Congreso Internacional de Lingüística de Corpus CILC 2014*. Las Palmas de Gran Canaria, 22-24 de mayo de 2014.
   [Visión general sobre el corpus MultiMedica.]