III CONGRESO INTERNACIONAL DE LINGÜÍSTICA DE CORPUS

DEPARTAMENTO DE LINGÜÍSTICA APLICADA
UNIVERSIDAD POLITÉCNICA DE VALENCIA
7-9 DE ABRIL DE 2011

LAS TECNOLOGÍAS DE LA INFORMACIÓN Y LAS COMUNICACIONES: PRESENTE Y FUTURO EN EL ANÁLISIS DE CÓRPORA

DEPARTAMENTO DE LINGÜÍSTICA APLICADA. EDIFICIO 4P
AVINGUDA TARONGERS, S/N 46022 VALENCIA
TEL. +963877530
FAX. +96377539
E-MAIL: cilc2011@upvnet.upv.es
WEB: WWW.CILC2011.UPV.ES

CONTENIDOS

BIENVENIDA

CILC 2011- VALENCIA

El Comité Organizador del III Congreso Internacional de la Asociación Española de Lingüística de Corpus (AELINCO) les da la bienvenida y agradece tanto a los ponentes como a los asistentes su participación en este encuentro anual de la Asociación. La Universidad Politécnica de Valencia, así como el departamento de Lingüística Aplicada han acogido este congreso con entusiasmo, recibiendo apoyo y ánimo para esta iniciativa.

Estamos muy satisfechos por la gran cantidad de propuestas recibidas tanto nacionales como internacionales, lo cual evidencia la importancia de este tipo de debates anuales sobre aspectos específicos de la Lingüística. En línea con el objetivo específico de AELINCO y de los congresos anteriores, la tercera edición del Congreso Internacional de Lingüística de Corpus se centra en la difusión de investigaciones desarrolladas en el marco de la Lingüística de Corpus y da cabida a estudios sobre distintos aspectos y aplicaciones del lenguaje natural o las lenguas particulares basados en el análisis de córpora mediante las herramientas ofrecidas por las tecnologías de la información y de las comunicaciones (TICs).

Esperamos que esta edición del congreso de la Asociación sea del agrado de todos los participantes, que disfruten del intercambio de investigaciones y proyectos que se presentan en los nueve paneles temáticos del congreso, así como de las ponencias plenarias.

Por supuesto, todo nuestro esfuerzo no obtendría su fruto sin la valiosa ayuda de todos aquellos que han participado en la organización del congreso y de las entidades financiadoras de este evento.

Benvinguts i benvingudes a València!

Esperamos que disfrutéis de vuestra estancia en Valencia.

EL COMITÉ ORGANIZADOR

CILC 2011

WELCOME

CILC 2011- VALENCIA

The organizing committee of the III International Congress of the Spanish Association of Corpus Linguistics (AELINCO) wishes you a warm welcome to Valencia and would like to thank all the speakers and attendees who are taking part in this annual meeting of the Association. Both the Universidad Politécnica de Valencia, and the Department of Applied Linguistics were enthusiastic about holding the congress here, and we are grateful for the support and encouragement given in order to bring this about.

We are delighted with the number of proposals submitted from within Spain itself, and from all over the world, which shows how relevant these annual events are in order to promote discussion and to reflect on specific aspects of studies in Linguistics. In line with the specific aims of AELINCO and previous Conferences, the third edition of the International Conference on Corpus Linguistics focuses on the dissemination of research conducted within the framework of Corpus Linguistics, including different aspects of natural language processing and corpus analysis using the different tools which have been developed in the field of Information and Communication Technologies (ICTs) for the study of specific languages and genre.

We sincerely hope that the present edition of the AELINCO congress will be a success, and that the participants enjoy having the opportunity to exchange ideas and inform each other about different research projects in the nine thematic panels and the plenary sessions.

Lastly, we would like to thank all those who have participated in the organization of the congress and the different sponsors, without whose help and finance the event would not have been possible.

Benvinguts i Benvingudes a València!

Welcome, and enjoy your stay in Valencia!

The Organizing Committee

CILC 2011

COMITÉ ORGANIZADOR/ORGANIZING COMMITTEE

COORDINACIÓN

María Luisa Carrió Pastor

SECRETARÍA ACADÉMICA

Ana Botella Trelis

Miguel Ángel Candel Mora

Luz Gil Salom

Penny MacDonald Lightbound

Carmen Soler Monreal

Keith Stuart

PANELES Y DIRECTORES/PANELS AND DIRECTORS

1. **Diseño, elaboración y tipología de corpus**
   Francisco Alonso Almeida
   Universidad de Las Palmas de Gran Canaria
   e-mail: falonso@dfm.ulpgc.es

2. **Discurso, análisis literario y corpus**
   José Luis Oncins
   Universidad de Cáceres
   e-mail: oncins@unex.es

3. **Gramática basada en corpus**
   Javier Pérez Guerra
   Facultade de Filoloxía e Tradución
   e-mail: jperez@uvigo.es

4. **Lexicología y lexicografía basadas en corpus**
   Pedro Fuertes Olivera
   Universidad de Valladolid
   e-mail: pedro@tita.emp.uva.es

5. **Corpus, estudios contrastivos y traducción**
   M. de los Ángeles Gómez
   Universidad de Santiago de Compostela
   e-mail: mdelosangeles.gomez@usc.es

6. **Variación lingüística y corpus**
   María José López Couso
   Universidade de Santiago de Compostela
   e-mail: mjlopez.couso@usc.es

7. **Lingüística computacional basada en corpus**
   Carlos Subirats
   International Computer Science Institute
   e-mail: carlos.subirats@gmail.com

8. **Corpus, adquisición y enseñanza de lenguas**
   Raquel Criado Sánchez
   Universidad de Murcia
   e-mail: rcriado@um.es

9. **Usos y aplicaciones específicas de la lingüística de corpus**
   Isabel de la Cruz Cabanillas
   Universidad de Alcalá
   e-mail: isabel.cruz@uah.es

JUNTA DIRECTIVA DE AELINCO/EXECUTIVE BOARD

Presidente
Aquilino Sánchez Pérez
Universidad de Murcia
e-mail: asanchez@um.es

Vicepresidente
Pascual Cantos Gómez
Universidad de Murcia
e-mail: pcantos@um.es

Secretario
Moisés Almela Sánchez
Universidad de Murcia
e-mail: moisesal@um.es

Tesorera
Nila Vázquez González
Universidad de Murcia
e-mail: nilavg@um.es

Vocal 1
Marisa Carrió Pastor
Universidad Politécnica de Valencia
e-mail: lcarrio@idm.upv.es

Vocal 2
Isabel Moskovich
Universidad de La Coruña
e-mail: imoskovich@udc.es

AGRADECIMIENTOS

Universidad Politécnica de Valencia

Vicerrectorado de Investigación

Centro de Formación Permanente

Departamento de Lingüística Aplicada

Generalitat Valenciana

Ministerio de Ciencia e Innovación

Editorial MacMillan

Editorial Routledge

Editorial Pearson Longman

Garnet Education

PLANOS DE SITUACIÓN/ GETTING HERE

VALENCIA- UPV

CÓMO LLEGAR/ GETTING HERE

DIRECCIÓN/ADDRESS:

UNIVERSIDAD POLITÉCNICA DE VALENCIA

DEPARTAMENTO DE LINGÜÍSTICA APLICADA. EDIFICIO 4P

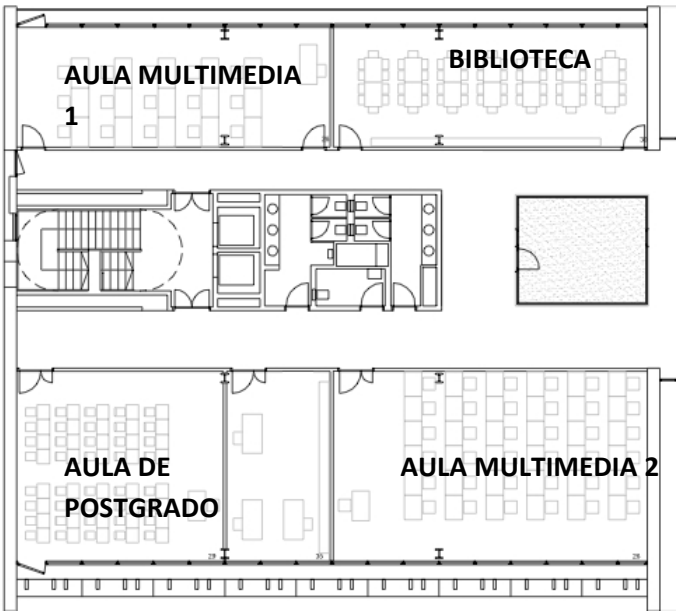
AVINGUDA TARONGERS

(ENFRENTE PARADA METRO-CARRASCA/EDIFICIO CONTIGUO A ETS DE TELECOMUNICACIONES)

HTTP://WWW.UPV.ES/PLANO/PLANO_UPVC.HTML

SITUACIÓN DE LAS SALAS DE PRESENTACIÓN/ROOM PLAN

DEPARTAMENTO DE LINGÜÍSTICA APLICADA. EDIFICIO 4P/BUILDING 4P



PLANTA 2

PLANTA 3

***PARA ACCEDER AL SALÓN DE ACTOS DE LA ETS DE TELECOMUNICACIONES SE HA DE IR A LA 2ª PLANTA, CRUZAR POR LA PASARELA DE COMUNICACIÓN DE LOS DOS EDIFICIOS Y SUBIR POR LAS ESCALERAS A LA 3ª PLANTA. SE COLOCARÁN INDICACIONES EN LA 2ª PLANTA DEL DEPARTAMENTO DE LINGÜÍSTICA APLICADA PARA FACILITAR EL ACCESO.**

INFORMACIÓN PARA LOS CONGRESISTAS

- **El material del Congreso se recogerá en la 3ª planta del Departamento de Lingüística Aplicada los días 7 y 8 de abril. Los congresistas que deseen hacerlo el día 9 de abril, deberán contactar con la organización del Congreso.**

- **Los congresistas dispondrán de 15 minutos de exposición y al final de cada sesión paralela se abrirá un turno de 5 minutos para el debate.**

- **Los congresistas han de consultar el tablón de anuncios del Congreso, situado en el 2º piso del Departamento de Lingüística Aplicada, para conocer las posibles incidencias y cambios en el Programa.**

- **Los pósters se expondrán en el pasillo de la 2º planta del Departamento de Lingüística Aplicada de 16.00 a 18.00 el viernes 8 de abril.**

- **Los congresistas dispondrán de conexión wi-fi en todo el edificio y podrán acceder mediante la clave que se les entregará junto con la documentación del congreso. Así mismo, también podrán acceder a Internet en el Aula Multimedia 2 (2º piso) cuando no se estén realizando sesiones paralelas.**

- **Todos los actos son de libre acceso excepto la cena de gala, para la cual se ha de realizar reserva (véase la web del congreso www.cilc2011.upv.es).**

- **Únicamente se entregará el certificado de participación a los ponentes que hayan pagado la cuota de inscripción y hayan presentado su ponencia.**

- **Aquellos ponentes que deseen publicar sus ponencias en las actas del congreso han de atenerse a las normas de publicación que pueden ~~ver~~ encontrar en la web del congreso y enviar su artículo a cilc2011@upvnet.upv.es hasta el 8 de mayo de 2011. Se seleccionarán artículos entre los recibidos para una publicación en una editorial internacional.**

INFORMATION FOR CONGRESS ATTENDEES

- **The Congress folders can be picked up in the Department of Applied Linguistics (3rd floor), on the 7th and 8th April. Any late arrivals on the 9th April should get in touch with the congress organisers for their documentation.**

- **Speakers will have 15 minutes for their communication and at the end of each panel session they will have 5 minutes each for questions.**

- **There will be a Congress notice board on the 2nd floor of the building indicating any last minute changes or any other alterations to the programme.**

- **The posters will be on show in the corridor on the 2nd floor of the Applied Linguistics Department from 16.00 to 18.00 on Friday 8th April.**

- **Attendees will have wi-fi access in the whole of the building and the password can be found in the Congress documents folder. Internet access is also available in the Aula Multimedia 2 (2nd floor) when this is free.**

- **All Congress events are open to all attendees, except the Gala dinner, which must be booked in advance (Congress website: www.cilc2011.upv.es).**

- **- Certificates will only be given to presenters who have paid the conference fee and have presented their paper.**

- **- Those speakers who wish to publish their papers in the conference proceedings must follow the style guidelines for publication that can be found on the conference website and send their article to cilc2011@upvnet.upv.es by May 8th, 2011. Selected articles amongst those received will be published in a special edition of an international publisher.**

PROGRAMA DEL CONGRESO CILC 2011

Jueves, 7 de abril de 2011

09.00-10.00 Entrega de documentación
Departamento de Lingüística Aplicada. Edificio 4P, 3º piso.

10.00-10.30 Acto de Inauguración por el Excmo. y Magfco. Rector de la Universidad Politécnica de Valencia
(Salón de Actos de la ETS Telecomunicaciones. 3er piso)

10.30-11.30 Conferencia Inaugural: Prof. Mike Scott (University of Aston)
Investigating Patterns
(Salón de Actos. ETS de Telecomunicaciones. 3er piso)

11.30-12.00 Descanso - Café en el Departamento de Lingüística Aplicada. Edificio 4P, 2º piso

12.00-14.00 **SESIONES PARALELAS I**
Panel 1: Diseño, elaboración y tipología de corpus (Aula multimedia 1, 2º piso)
Panel 2: Discurso, análisis literario y corpus (Biblioteca, 2º piso)
Panel 4: Lexicología y lexicografía basadas en corpus (Aula de Posgrado, 2ª piso)
Panel 5: Corpus, estudios contrastivos y traducción (Aula multimedia 2, 2º piso)
Panel 8: Corpus, adquisición y enseñanza de lenguas (Aula multimedia 3, 3er piso)

14.00-16.00 Descanso

16.00-17.00 Conferencia Plenaria: Prof. Javier Martín Arista (Universidad de La Rioja)
Uso de corpora lexicográficos y textuales para la elaboración de una base de datos léxica
(Salón de Grados. Departamento de Lingüística Aplicada. 3er piso)

17.00-17.30 Descanso - Café en el Departamento de Lingüística Aplicada. Edificio 4P, 2º piso

17.30-19.30 **SESIONES PARALELAS II**
Panel 1: Diseño, elaboración y tipología de corpus (Aula multimedia 1, 2º piso)
Panel 3: Gramática basada en corpus (Aula multimedia 3, 3er piso)
Panel 4: Lexicología y lexicografía basadas en corpus (Salón de grados, 3er piso)
Panel 5: Corpus, estudios contrastivos y traducción (Aula multimedia 2, 2º piso)
Panel 6: Variación lingüística y corpus (Aula de Posgrado, 2º piso)
Panel 7: Lingüística computacional basada en corpus (Biblioteca, 2º piso)

20.30 Recepción de bienvenida. Hotel Astoria. Salón-Terraza, 9ª planta. Plaza Rodrigo Botet, 5.

Viernes, 8 de abril de 2011

10.00-11.30 **SESIONES PARALELAS III**
Panel 1: Diseño, elaboración y tipología de corpus (Aula multimedia 1, 2º piso)
Panel 2: Discurso, análisis literario y corpus (Biblioteca, 2º piso)
Panel 4: Lexicología y lexicografía basadas en corpus (Salón de grados, 3er piso)
Panel 5: Corpus, estudios contrastivos y traducción (Aula multimedia 2, 2º piso)
Panel 8: Corpus, adquisición y enseñanza de lenguas (Aula multimedia 3, 3er piso)
Panel 9: Usos y aplicaciones específicas de la lingüística de corpus (Aula de Posgrado, 2º piso)

11.30-12.00 Descanso - Café en el Departamento de Lingüística Aplicada. Edificio 4P, 2º piso

12.00-13.00 Conferencia Plenaria: Profa. Susan Hunston (University of Birmingham)
Patterns and Evaluative Meaning
(Salón de Actos. ETS de Telecomunicaciones. 3er piso)

13.00-14.00 **SESIONES PARALELAS IV**

Panel 4: Lexicología y lexicografía basadas en corpus (Salón de grados, 3er piso)
Panel 5: Corpus, estudios contrastivos y traducción (Aula multimedia 1, 2º piso)
Panel 6: Variación lingüística y corpus (Aula de Posgrado, 2º piso)
Panel 7: Lingüística computacional basada en corpus (Aula multimedia 2, 2º piso)
Panel 8: Corpus, adquisición y enseñanza de lenguas (Biblioteca, 2º piso)

14.00-16.00 Descanso

16.00-18.00 **SESIONES PARALELAS V**
Panel 1: Diseño, elaboración y tipología de corpus (Aula multimedia 1, 2º piso)
Panel 2: Discurso, análisis literario y corpus (Aula de posgrado, 2º piso)
Panel 3: Gramática basada en corpus (Aula multimedia 3, 3er piso)
Panel 4: Lexicología y lexicografía basadas en corpus (Salón de grados, 3er piso)
Panel 5: Corpus, estudios contrastivos y traducción (Aula multimedia 2, 2º piso)
Panel 8: Corpus, adquisición y enseñanza de lenguas (Biblioteca, 2º piso)
Exposición de pósters- 2º piso.

18.00-18.30 Descanso - Café en el Departamento de Lingüística Aplicada. Edificio 4P, 2º piso

18.30-19.30 Conferencia Plenaria: Prof. Mike O'Donnell (Universidad Autónoma de Madrid)
Using learner corpora to redesign university-level ESL education
(Salón de Grados. Departamento de Lingüística Aplicada. 3er piso)

19.30-21.00 **ASAMBLEA GENERAL DE SOCIOS DE AELINCO**
(Salón de Grados. Departamento de Lingüística Aplicada. 3er piso)

21.30 Cena de Gala. Hotel Westin. Salón Exposición. (C/ Amadeo de Saboya, 16)
(Véase inscripción en web)

Sábado, 9 de abril de 2011

10.00-11.30 **SESIONES PARALELAS VI**
Panel 2: Discurso, análisis literario y corpus (Biblioteca, 2º piso)
Panel 5: Corpus, estudios contrastivos y traducción (Aula multimedia 1, 2º piso)
Panel 6: Variación lingüística y corpus (Aula de Posgrado, 2º piso)
Panel 8: Corpus, adquisición y enseñanza de lenguas (Biblioteca, 2º piso)
Panel 9: Usos y aplicaciones específicas de la lingüística de corpus (Aula Multimedia 2, 2º piso)

11.30-12.00 Descanso - Café en el Departamento de Lingüística Aplicada. Edificio 4P, 2º piso

12.00-13.00 **SESIONES PARALELAS VII**
Panel 2: Discurso, análisis literario y corpus (Biblioteca, 2º piso)
Panel 4: Lexicología y lexicografía basadas en corpus (Salón de grados, 3er piso)
Panel 6: Variación lingüística y corpus (Aula Multimedia 1, 2º piso)
Panel 8: Corpus, adquisición y enseñanza de lenguas (Aula Multimedia 3, 3er piso)
Panel 9: Usos y aplicaciones específicas de la lingüística de corpus (Aula de Posgrado, 2º piso)

13.00-14.15 Conferencia de Clausura: Prof. Antonio Briz (Universitat de València)
Los corpus orales del español: la cualidad y la cantidad de los datos
(Salón de Grados. Departamento de Lingüística Aplicada. 3er piso)

14.15 Acto de Clausura
(Salón de Grados. Departamento de Lingüística Aplicada, 3er piso)
Vino de Honor
(Entrada principal del edificio 4P, Departamento de Lingüística Aplicada)

CONFERENCE PROGRAMME

Thursday, 7 April, 2011

09.00-10.00 Registration
Department of Applied Linguistics, Building 4P, 4th Floor.

10.00-10.30 Opening Ceremony: Rector, Universidad Politécnica de Valencia
(Salón de Actos de la ETS Telecomunicaciones, 3rd Floor)

10.30-11.30 Inaugural Conference: Prof. Mike Scott (University of Aston)
Investigating Patterns
(Salón de Actos. ETS de Telecomunicaciones, 3rd Floor)

11.30-12.00 Coffee Break – Department of Applied Linguistics. Building 4P, 2nd Floor

12.00-14.00 **PARALLEL SESSION I**
Panel 1: Corpus design, development and typology (Aula multimedia 1, 2nd Floor)
Panel 2: Discourse, literary analysis and corpora (Biblioteca, 2nd Floor)
Panel 4: Corpus-based lexicology and lexicography (Salón de grados, 3rd Floor)
Panel 5: Corpora, contrastive studies and translation (Aula multimedia 2, 2nd Floor)
Panel 8: Corpora, language acquisition and teaching (Aula multimedia 3, 3rd Floor)

14.00-16.00 Lunch

16.00-17.00 Plenary Conference: Prof. Javier Martín Arista (Universidad de La Rioja)
Use of lexicographic and textual corpora for the development of a lexical database
(Salón de Grados, Department of Applied Linguistics, 3rd Floor)

17.00-17.30 Coffee Break - Department of Applied Linguistics. Building 4P, 2nd Floor

17.30-19.30 **PARALLEL SESSION II**
Panel 1: Corpus design, development and typology (Aula multimedia 1, 2nd Floor)
Panel 3: Corpus-based grammatical studies (Aula multimedia 3, 2nd Floor)
Panel 4: Corpus-based lexicology and lexicography (Salón de grados, 3rd Floor)
Panel 5: Corpora, contrastive studies and translation (Aula multimedia 2, 2nd Floor)
Panel 6: Linguistic variation and corpus (Aula de Posgrado, 2nd Floor)
Panel 7: Corpus-based computational linguistics (Biblioteca, 2nd Floor)

20.30 Welcome Reception. Hotel Astoria, Salón-Terraza, 9th Floor. Plaza Rodrigo Botet, 5.

Friday, April 8, 2011

10.00-11.30 **PARALLEL SESSION III**
Panel 1: Corpus design, development and typology (Aula multimedia 1, 2nd Floor)
Panel 2: Discourse, literary analysis and corpora (Biblioteca, 2nd Floor)
Panel 4: Corpus-based lexicology and lexicography (Salón de grados, 3rd Floor)
Panel 5: Corpora, contrastive studies and translation (Aula multimedia 2, 2nd Floor)
Panel 8: Corpora, language acquisition and teaching (Aula multimedia 3, 3º Floor)
Panel 9: Corpus linguistics: Uses and specific applications (Aula de Posgrado, 2nd Floor)

11.30-12.00 Coffee Break - Department of Applied Linguistics, Building 4P, 2nd Floor

12.00-13.00 Plenary Conference: Prof. Susan Hunston (University of Birmingham)
Patterns and Evaluative Meaning
(Salón de Actos. ETS de Telecomunicaciones. 3rd Floor)

13.00-14.00 **PARALLEL SESSION IV**
Panel 4: Corpus-based lexicology and lexicography (Salón de grados, 3rd Floor)
Panel 5: Corpora, contrastive studies and translation (Aula multimedia 1, 2nd Floor)
Panel 6: Linguistic variation and corpus (Aula de Posgrado, 2nd Floor)
Panel 7: Corpus-based computational linguistics (Aula multimedia 2, 2nd Floor)
Panel 8: Corpora, language acquisition and teaching (Biblioteca, 2nd Floor)

14.00-16.00 Lunch

16.00-18.00 **PARALLEL SESSION V**
Panel 1: Corpus design, development and typology (Aula multimedia 1, 2nd Floor)
Panel 2: Discourse, literary analysis and corpora (Biblioteca, 2nd Floor)
Panel 3: Corpus-based grammatical studies (Aula multimedia 3, 2nd Floor)
Panel 4: Corpus-based lexicology and lexicography (Salón de grados, 3rd Floor)
Panel 5: Corpora, contrastive studies and translation (Aula multimedia 2, 2nd Floor)
Panel 8: Corpora, language acquisition and teaching (Biblioteca, 2nd Floor)
Poster Exhibition - 2nd Floor.

18.00-18.30 Coffee Break - Department of Applied Linguistics, Building 4P, 2nd Floor

18.30-19.30 Plenary Conference: Prof. Michael O'Donnell (Universidad Autónoma de Madrid)
Using learner corpora to redesign university-level ESL education
(Salón de Grados. Department of Applied Linguistics, 3rd Floor)

19.30-21.00 **GENERAL ASSEMBLY OF AELINCO MEMBERS**
(Salón de Grados. Department of Applied Linguistics, 3rd Floor)

21.30 Gala Dinner. Hotel Westin. Salón Exposición. (C/ Amadeo de Saboya, 16)
(See web registration)

Saturday, April 9, 2011

10.00-11.30 **PARALLEL SESSION VI**
Panel 2: Discourse, literary analysis and corpora (Biblioteca, 2nd Floor)
Panel 5: Corpora, contrastive studies and translation (Aula multimedia 1, 2nd Floor)
Panel 6: Linguistic variation and corpus (Aula de Posgrado, 2nd Floor)
Panel 8: Corpora, language acquisition and teaching (Biblioteca, 2nd Floor)
Panel 9: Corpus linguistics: Uses and specific applications (Aula de Posgrado, 2nd Floor)

11.30-12.00 Coffee Break - Department of Applied Linguistics, Building 4P, 2nd Floor

12.00-13.00 **PARALLEL SESSION VII**
Panel 2: Discourse, literary analysis and corpora (Biblioteca, 2nd Floor)
Panel 4: Corpus-based lexicology and lexicography (Salón de grados, 3rd Floor)
Panel 6: Linguistic variation and corpus (Aula de Posgrado, 2nd Floor)
Panel 8: Corpora, language acquisition and teaching (Biblioteca, 2nd Floor)
Panel 9: Corpus linguistics: Uses and specific applications (Aula de Posgrado, 2nd Floor)

13.00-14.15 Conferencia de Clausura: Prof. Antonio Briz (Universitat de València)
Spanish oral corpora: data quantity and quality
(Salón de Grados. Department of Applied Linguistics, 3rd Floor)

14.15 Closing Ceremony
(Salón de Grados. Department of Applied Linguistics, 3rd Floor)
Wine Reception
(Entrance to Department of Applied Linguistics, Building 4P)

SESIONES PARALELAS DE PRESENTACIÓN DE PONENCIAS/PARALLEL SESSIONS

Jueves/Thursday, 7 de abril de 2011
12.00-14.00

Aula multimedia 1, 2º piso/floor

Panel 1: Diseño, elaboración y tipología de corpus (Dr. Francisco Alonso Almeida)
Panel 1: Corpus design, development and typology

Jesús Romero-Trillo, Silvia Riesco-Bernier, Karina Vidal, Belén Díez-Bedmar, Teresa Gerdes, Anna Gladkova, Elizabeth Lenn and Tíscar Espigares
CORPUS OF LANGUAGE AND NATURE (CLAN-PROJECT): THE REPRESENTATION OF LANDSCAPE UNIVERSALS IN LANGUAGE

Laura Ramírez Polo
MATVA: A DATABASE OF ENGLISH TELEVISION COMMERCIALS FOR THE STUDY OF PRAGMATIC-COGNITIVE EFFECTS OF PARALINGUISTIC AND EXTRALINGUISTIC ELEMENTS ON THE AUDIENCE OF ENGLISH TV ADS

Marta Conejero, Asunción Jaime and Debra Westall
NIP & TUCK: A CORPUS-BASED QUALITATIVE TYPOLOGY FOR CONCISION IN SCIENTIFIC WRITING

Joseba Ezeiza and Agurtzane Elordui
HERRAMIENTAS Y CRITERIOS PARA LA CREACIÓN DE UN BANCO DE CONOCIMIENTO SOBRE LOS USOS DEL LENGUAJE EN LA RED

Biblioteca, 2º piso/floor

Panel 2: Discurso, análisis literario y corpus (Dr. José Luis Oncins)
Panel 2: Discourse, literary analysis and corpora

Hanna Skorczynska
METAPHOR IDENTIFICATION IN CORPORA: THE CASE OF 'AS' IN A BUSINESS PERIODICAL CORPUS

David Brown and Laura Aull
"TOUGH GUYS" AND "CATFIGHT CRAZY": A CORPUS-BASED ANALYSIS OF GENDER REPRESENTATIONS IN SPORTS REPORTAGE

Ángela Almela and Gema Alcaraz
MEASURING WILDE'S STYLE: AN APPLICATION OF COMPUTER STYLOMETRY TO A LITERARY CORPUS

Salón de Grados, 3er piso /floor

Panel 4: Lexicología y lexicografía basadas en corpus (Dr. Pedro Fuertes Olivera)
Panel 4: Corpus-based lexicology and lexicography

Beatriz Sánchez Cárdenas and Pamela Faber Benítez

LA PROTOTIPICIDAD DE LOS ARGUMENTOS VERBALES COMO FACTOR DELIMITADOR DE LA ESTRUCTURA JERÁRQUICA DE UN DOMINIO LÉXICO

Mojca Kompara
IS AUTOMATIC PRODUCTION OF DICTIONARY ENTRIES IN THE FIRST SLOVENE ONLINE DICTIONARY OF ABBREVIATIONS SLOVARČEK KRAJŠAV POSSIBLE?

Serge Potemkin
SENTIMENT EXTRACTION FROM THE BILINGUAL CORPUS

Belén López Arroyo and Martín Fernández Antolín
CORPUS BASED APPLICATIONS: DEFINING A BILINGUAL LEXICOGRAPHICAL AND PHRASEOLOGICAL WORK ON WINE TASTING NOTES

Aula Multimedia 2, 2º piso/floor

Panel 5: Corpus, estudios contrastivos y traducción (Dra. Mª Ángeles Gómez)
Panel 5: Corpora, contrastive studies and translation

Francisco Alonso-Almeida and Ivalla Ortega-Barrera
EVIDENTIALITY AND EPISTEMIC MODALITY IN ENGLISH AND SPANISH LEGAL SCIENTIFIC DISCOURSE: A CORPUS-BASED STUDY

Taner Karakoc
CORPUS OF TURKISH CULTURE-SPECIFIC ITEMS AS REPRESENTATIVES THROUGH TRANSLATION IN ISTANBUL 2010 EUROPEAN CAPITAL OF CULTURE ACTIVITIES

Francisco Gonzálvez-García
THE GRAMMAR-DISCOURSE INTERFACE REVISITED WITHIN CONTRASTIVE CONSTRUCTION GRAMMAR: THE CASE OF FOCUS CONSTRUCTIONS IN ENGLISH AND SPANISH

Noelia Ramon
'WELL' IN SPANISH TRANSLATIONS: EVIDENCE FROM THE P-ACTRESS PARALLEL CORPUS

Mariana Orozco-Jutorán
EL USO INTEGRADO DE CORPUS Y MEMORIAS DE TRADUCCIÓN: CÓMO SACAR EL MÁXIMO PARTIDO DE LAS NUEVAS TECNOLOGÍAS PARA LA TRADUCCIÓN JURÍDICA

Patrick Goethals
DEMONSTRATIVE MODIFIERS AND DEFINITE ARTICLES IN TRANSLATION: A CONTRASTIVE PERSPECTIVE

Aula multimedia 3, 3er piso/floor

Panel 8: Corpus, adquisición y enseñanza de lenguas (Dra. Raquel Criado Sánchez)
Panel 8: Corpora, language acquisition and teaching

Daniela Gil-Salom
LA ADQUISICIÓN DE ALEMÁN COMO LENGUA EXTRANJERA. UNA APORTACIÓN BASADA EN CORPUS DE APRENDICES

Sánchez Aquilino, Cantos Pascual and Criado-Sánchez Raquel
CORPORA-BASED FREQUENCY LISTS, READABILITY INDEX AND ELT TEXTBOOKS

Gema Alcaraz-Mármol and Lourdes Cerezo-García

SPECIFIC FREQUENCY AND ITS ROLE IN FOREIGN LANGUAGE VOCABULARY ACQUISITION

Su-han Cheng and Jeng-yih Hsu
A CORPUS-BASED STUDY OF THE VOCABULARY USE IN AN ENGLISH NEWSPAPER

Jueves/Thursday, 7 de abril de 2011
17.30-19.30

Aula multimedia 1, 2º piso/floor
Panel 1: Diseño, elaboración y tipología de corpus (Dr. Francisco Alonso Almeida)
Panel 1: Corpus design, development and typology

Isabel Duran
CRITERIOS ESPECÍFICOS PARA LA ELABORACIÓN Y DISEÑO DE LOS CORPUS ESPECIALIZADOS PARA LA TERMINOGRAFÍA

Tanja Wissik
COMPILING SPECIALIZED CORPORA ACROSS LANGUAGE VARIETIES AND WORKING WITH THEM

Karlheinz Moerth, Niku Dorostkar and Alexander Preisinger
GLEANING MICRO-CORPORA FROM THE INTERNET: INTEGRATING HETEROGENEOUS DATA INTO EXISTING CORPUS INFRASTRUCTURES

Hanna Hedeland
INTERACTION OF TECHNOLOGY AND METHODOLOGY IN BUILDING AND SHARING AN ANNOTATED LEARNER CORPUS OF SPOKEN GERMAN

Dionysis Goutsos, Constantin Potagas, Dimitris Kasselimis, Maria Varkanitsa & Ioannis Evdokimidis
THE CORPUS OF GREEK APHASIC SPEECH: DESIGN AND COMPILATION

Lautenai Antonio Bartholamei Junior
PEPCO: DESIGNING A PARALLEL AND COMPARABLE TRANSLATIONAL CORPUS IN BRAZIL

Gunta Nešpore, Lauma Pretkalniņa, Baiba Saulīte and Kristīne Levāne-Petrova
TOWARDS A LATVIAN TREEBANK

Aula multimedia 3, 3er piso /floor
Panel 3: Gramática basada en corpus (Dr. Javier Pérez Guerra)
*Panel 3: Corpus-based grammatical studi*es

Peter Bouda
LANGUAGE DOCUMENTATION CORPORA IN DESCRIPTIVE LINGUISTICS

João Henrique Rettore-Totaro
MENSURACIÓN DE LA VARIABILIDAD ESTRUCTURAL EN CORPORA ROMÁNICOS MEDIEVALES Y MODERNOS

Pau Giménez, Joan Costa, Aina Labèrnia and Àlex Alsina
EL PROYECTO DELADI: EVALUACIÓN DEL CONOCIMIENTO Y USO DE LOS PRONOMBRES RELATIVOS EN CATALÁN

Mariya Khudyakova
POSSESSOR NPS AND REFERENTIAL CHOICE IN ENGLISH BUSINESS PROSE (A CORPUS RESEARCH)

Lien De Vos
THE USE OF GENDER-MARKED PRONOUNS IN DUTCH: GRAMMATICAL VERSUS CONCEPTUAL GENDER

Salón de Grados, 3er piso/floor

Panel 4: Lexicología y lexicografía basadas en corpus (Dr. Pedro Fuertes Olivera)
Panel 4: Corpus-based lexicology and lexicography

Irene Renau y Rogelio Nazar
ANÁLISIS CUANTITATIVO DEL USO REAL DE LOS VERBOS PRONOMINALES ESTRICTOS DEL CASTELLANO UTILIZANDO UN CORPUS DIACRÓNICO (GOOGLE BOOKS)

Julia Sanmartín Sáez and Nuria Edo Marzá
ANÁLISIS DEL CONCEPTO 'HABITACIÓN' EN UN CORPUS BILINGÜE ESPAÑOL-INGLÉS DE PÁGINAS ELECTRÓNICAS DE PROMOCIÓN HOTELERA

Elena Quintana Toledo and Margarita Esther Sánchez Cuervo
AN APPROACH TO TYPES OF MODALITY IN THE INTRODUCTION AND THE CONCLUSION SECTIONS OF COMPUTING RESEARCH ARTICLES

Carmen Ávila Martín and Ramón Martí Solano
EL ANÁLISIS DISCURSIVO DE LA VIOLENCIA A TRAVÉS DE UN CORPUS ESPECÍFICO

Isabel Marcelino, Gaël Dias, João Casteleiro and José Martinez-De-Oliveira
SEMI-AUTOMATIC CONSTRUCTION OF THE UNIFIED MEDICAL LEXICON FOR PORTUGUESE

Aula multimedia 2, 2º piso /floor

Panel 5: Corpus, estudios contrastivos y traducción (Dra. Mª Ángeles Gómez)
Panel 5: Corpora, contrastive studies and translation

Miguel Angel Candel-Mora and Chelo Vargas Sierra
ANÁLISIS DE LA PRODUCCIÓN INVESTIGADORA EN LINGÜÍSTICA DE CORPUS APLICADA A LA TRADUCCIÓN

Lourdes Juncal
A CONTRASTIVE STUDY OF ADVERBS OF CERTAINTY AS DISCOURSE MARKERS IN SPOKEN ENGLISH AND SPANISH

Maria Josep Cuenca and Josep Ribera
DEICTIC NEUTRALIZATION AND OVERMARKING IN TRANSLATING FICTION (ENGLISH-CATALAN)

Belén López Arroyo
WRITING COMPUTERIZED ABSTRACTS: APPLICATIONS FROM A CORPUS-BASED STUDY

Ángela Almela and Samuel Gracia

EL GUIÓN CINEMATOGRÁFICO COMO CORPUS: UN ESTUDIO CONTRASTIVO ENTRE EL ESPAÑOL CASTIZO DE ALMODÓVAR Y SU TRADUCCIÓN AL INGLÉS

Daniel Gallego-Hernández and Ramesh Krishnamurthy
COMENEGO (CORPUS MULTILINGÜE DE ECONOMÍA Y NEGOCIOS) VS. METODOLOGÍAS WEB AS/FOR CORPUS APLICADAS A LA PRÁCTICA DE LA TRADUCCIÓN ECONÓMICA, COMERCIAL Y FINANCIERA

Aula de Posgrado, 2º piso/floor
Panel 6: Variación lingüística y corpus (Dra. María José López Couso)
Panel 6: Linguistic variation and corpus

Barry Pennock-Speck
VOICE-OVERS IN BRITISH TELEVISION ADS: A CORPUS ANALYSIS OF A WRITTEN-TO-BE-SPOKEN GENRE

Javier Ruano-Garcia
THE WORLD HAS GOT SOME HINT OF HER COUNTRY SPEECH: ON THE ENREGISTERMENT OF THE 'NORTHERN DIALECT'

Chris Culy, Verena Lyding and Henrik Dittmann
STRUCTURED PARALLEL COORDINATES: A VISUALIZATION FOR ANALYZING STRUCTURED LANGUAGE DATA

Gerold Schneider and Fabio Rinaldi
A DATA-DRIVEN APPROACH TO ALTERNATIONS BASED ON PROTEIN-PROTEIN INTERACTIONS

Fatima Faya Cerqueiro
REQUEST MARKERS IN DRAMA: DATA FROM THE CORPUS OF IRISH ENGLISH

Biblioteca, 2º piso/floor
Panel 7: Lingüística computacional basada en corpus (Dr. Carlos Subirats)
Panel 7: Corpus-based computational linguistics

Antonio Frías Delgado
ESTUDIO COMPARATIVO DE COLOCACIONES EN TEXTOS ORIGINALES Y EN SU TRADUCCIÓN

Irene Castellón, German Rigau, Salvador Climent, Marta Coll-Florit and Marina Lloberes
ANOTACIÓN SEMÁNTICA DEL CORPUS SENSEM

Marc Ortega Gil
ANÁLISIS LÉXICO DE UNIDADES LÉXICAS COMPUESTAS

Gotzon Aurrekoetxea
"CORPUSLEM" UNA HERRAMIENTA PARA LA CONVERSIÓN DE CORPUS TEXTUALES EN DATOS

Garazi Olaziregi, Francisco Javier Calle and Dolores Cuadra Fernández
COGNOS TOOLKIT: UN CONJUNTO DE HERRAMIENTAS PARA LA ANOTACIÓN LINGÜÍSTICA DE CORPUS

Viernes/Friday, 8 de abril de 2011
10.00-11.30

Aula multimedia 1, 2º piso/floor

Panel 1: Diseño, elaboración y tipología de corpus (Dr. Francisco Alonso Almeida)
Panel 1: Corpus design, development and typology

Miriam Seghiri
COMBITUR: ASPECTOS DE DISEÑO, COMPILACIÓN Y REPRESENTATIVIDAD DE UN CORPUS DE CONDICIONES GENERALES DE VIAJE COMBINADO

Ekaterina Tarpomanova, Svetlozara Leseva, Svetla Koeva, Borislav Rizov, Hristina Kukova, Tsvetana Dimitrova and Maria Todorova
DESIGN AND DEVELOPMENT OF THE BULGARIAN SENSE-ANNOTATED CORPUS

Paula Rodriguez-Puente
INTRODUCING THE CORPUS OF HISTORICAL ENGLISH LAW REPORTS: STRUCTURE AND COMPILATION TECHNIQUES

Heather Froehlich
ARE YOU A MAN? ON SEEING GENDER IN SHAKESPEARE

Biblioteca, 2º piso/floor

Panel 2: Discurso, análisis literario y corpus (Dr. José Luis Oncins)
Panel 2: Discourse, literary analysis and corpora

María Alcantud Díaz
VIOLENCE IN CHILDREN'S TALES: A SYSTEMIC CORPUS AND CRITICAL DISCOURSE ANALYSIS OF CINDERELLA

Kieran O'Halloran
ELECTRONIC DECONSTRUCTION OF AN ARGUMENT THROUGH ITS 'SUPPLEMENT': DERRIDA AND CORPUS LINGUISTIC METHOD

Georgia Fragaki
EVALUATIVE ADJECTIVES IN A CORPUS OF GREEK OPINION ARTICLES

Keith Stuart
A CORPUS ANALYSIS OF RHETORICAL STRATEGIES IN THE DISCOURSE OF CHOMSKY

Debra Westall
EL PAÍS NEWS REPORTS ON CHILDHOOD OBESITY: A TWELVE-MONTH CORPUS STUDY

Sergio Lobejón Santos
EL CORPUS TRACE, O CÓMO DISEÑAR UN CORPUS Y NO FRACASAR EN EL INTENTO

Salón de Grados, 3er piso/floor

Panel 4: Lexicología y lexicografía basadas en corpus (Dr. Pedro Fuertes Olivera)
Panel 4: Corpus-based lexicology and lexicography

Moisés Almela
FROM COLLOCATION TO INTER-COLLOCATION: DEVELOPING A DYNAMIC APPROACH TO COMBINATORIAL LEXICOGRAPHY

Raquel Vea
THE CORPUS PRODUCTIVITY OF OLD ENGLISH ADJECTIVAL COMPOUNDS WITH VERBAL BASE

Kornélia Papp
A CORPUS-BASED STUDY OF THE PROPERTY CONCEPTS KIS/KICSI 'SMALL' IN HUNGARIAN

Bernadette Borosi
CORPUS PARALELOS ALINEADOS: SEGMENTACIÓN TEXTUAL CON FINES LEXICOGRÁFICOS

Aula multimedia 2, 2º piso/floor

Panel 5: Corpus, estudios contrastivos y traducción (Dra. Mª Ángeles Gómez)
Panel 5: Corpora, contrastive studies and translation

Norsimah Mat Awal, Imran Ho-Abdullah and Intan Zainudin
A CORPUS-BASED STUDY ON THE LEXICO-GRAMMARTICAL DIVERGENCE IN MALAY TRANSLATED TEXT: AN ANALYSIS OF THE RELATIVE CLAUSE MARKER YANG

Ana Patricia García Varela
'WHEN POLICE ARRIVED AT THE SCENE' OR 'HAN VENIDO DOS POLICÍAS': ON THEME AND THEMATIC PROGRESSION IN NEWS REPORTS

Renata Enghels y Marlies Jansegers
HACIA UN ENFOQUE EMPÍRICO EN LA SEMÁNTICA: EL PAPEL DE LA TRADUCCIÓN. ESTUDIO CONTRASTIVO DEL VERBO SENTIR

Beatriz Rodríguez Arrizabalaga
THE BIRTH OF A NEW RESULTATIVE CONSTRUCTION IN SPANISH

Dámaso Izquierdo Alegría and Ramón González Ruiz
CORPUS PARALELOS Y ANÁLISIS DEL DISCURSO: PROPUESTAS DE EXPLOTACIÓN A PARTIR DEL ESTUDIO DE UN MECANISMO COHESIVO

Aula multimedia 3, 3er piso/floor

Panel 8: Corpus, adquisición y enseñanza de lenguas (Dra. Raquel Criado Sánchez)
Panel 8: Corpora, language acquisition and teaching

Joseba Ezeiza
PLATAFORMA GARALEX: INFRAESTRUCTURA TECNOLÓGICA PARA LA INVESTIGACIÓN Y LA DIDÁCTICA DE LENGUAJE DEL ÁMBITO DE LAS CIENCIAS JURÍDICAS

Natalia Judith Laso, Elisabet Comelles and Isabel Verdaguer
USING A CORPUS-BASED CLAUSE PATTERN DATABASE IN THE ENGLISH GRAMMAR CLASSROOM

María Belén Díez Bedmar

SPANISH STUDENTS' MAIN PROBLEMS WHEN WRITING THE ENGLISH EXAM IN THE UNIVERSITY ENTRANCE EXAMINATION: A LEARNER CORPUS-BASED ANALYSIS

Miguel Fuster Márquez and Begoña Clavel Arroitia
ENGLISH LANGUAGE TEACHING AND LEARNING IN TERTIARY EDUCATION: CORPUS CHOICE AND IMPLEMENTATION

Pansa Prommas and Kemtong Sinwongsuwat
A COMPARATIVE STUDY OF DISCOURSE CONNECTORS USED IN ARGUMENTATIVE COMPOSITIONS OF THAI EFL LEARNERS AND ENGLISH-NATIVE SPEAKERS

Aula de Posgrado, 2º piso/floor

Panel 9: Usos y aplicaciones específicas de la lingüística de corpus (Dra. Isabel de la Cruz Cabanillas)

Panel 9: Corpus linguistics: Uses and specific applications

Maria Luisa Carrio Pastor and Eva Mestre Mestre
THE USE OF CORPUS ANALYSIS TO MANAGE FOREIGN LANGUAGE ACQUISITION IN A BILINGUAL COMMUNITY

Pedro Alvarez Mosquera
TESTING THE EXCEPTION: AN ANALYSIS OF EMINEM'S LANGUAGE USES FROM A CORPUS-BASED APPROACH.

Rema Rossini, Fabio Tamburini and Andrea Zaninello
EXPLOITING CORPUS EVIDENCE FOR AUTOMATIC SENSE INDUCTION

David Brett and Antonio Pinna
LEXICAL BUNDLES IN US PRESIDENTIAL SPEECHES: A CORPUS-DRIVEN STUDY OF B. CLINTON'S, G.W. BUSH'S AND B. OBAMA'S ADDRESSES

Viernes/Friday, 8 de abril de 2011
13.00-14.00

Salón de Grados, 3er piso/floor

Panel 4: Lexicología y lexicografía basadas en corpus (Dr. Pedro Fuertes Olivera)

Panel 4: Corpus-based lexicology and lexicography

María Teresa Ortego
LA COMPILACIÓN DE DICOENVIRO EN ESPAÑOL (DICTIONNAIRE FONDAMENTAL DE L'ENVIRONNMENT)

Mojca Kompara, Ana Begus and Elena Sverko
COMBINED APPROACH TO MODERN LEXICOGRAPHIC TOOLS: THE CASE OF THE FIRST SLOVENE DICTIONARY OF TOURISM TERMINOLOGY

Araceli Alonso Campo
COLLOCATIONAL NETWORKS Y EL USO 'ESPECIALIZADO' Y 'GENERAL' DE LAS UNIDADES LÉXICAS: EL CASO DE AQUALEXIC

Garikoitz Knörr and Keith Stuart
THE SENSE AND SYNTAX OF 'SPEAK' AND 'TALK'

Aula multimedia 1, 2º piso/floor
Panel 5: Corpus, estudios contrastivos y traducción (Dra. Mª Ángeles Gómez)
Panel 5: Corpora, contrastive studies and translation

Maria Calzada Perez
ANÁLISIS CRÍTICOS DE DISCURSOS PARLAMENTARIOS EUROPEOS. DESDE LA TEXTURA AL CONTEXTO CON ECPC

Monica Palmerini and Serenella Zanotti
A CORPUS-BASED STUDY ON THE USE OF NARRATIVE IN ENGLISH AND SPANISH YOUTH CONVERSATIONS

Irina Keshabyan
A CONTRASTIVE STRUCTURAL ANALYSIS OF SHAKESPEARE'S HAMLET VERSUS SUMAROKOV'S GAMLET: A CORPUS-BASED APPROACH

José Manuel Martínez Martínez
¡HOUSTON, TENEMOS UN PROBLEMA... DE TRADUCCIÓN! ECPC Y TPC COMO HERRAMIENTAS DIDÁCTICAS PARA LA ENSEÑANZA/APRENDIZAJE DE LA TRADUCCIÓN

Aula de Posgrado, 2º piso/floor
Panel 6: Variación lingüística y corpus (Dra. María José López Couso)
Panel 6: Language variation and corpus

Carmen Soler-Monreal and Luz Gil-Salom
LITERATURE REVIEWS IN ENGLISH AND SPANISH PHD THESES: A CROSS-LANGUAGE STUDY

María José Luzón
DISCIPLINARY DIFFERENCES IN THE USE OF SUB-TECHNICAL NOUNS: A CORPUS-BASED STUDY

Mercedes Bengoechea and José Simón
FEMINIST LANGUAGE REFORM IN SPANISH ADVERTISING. A CORPUS-BASED RESEARCH

Aula Multimedia 2, 2º piso/floor
Panel 7: Lingüística computacional basada en corpus (Dr. Carlos Subirats)
Panel 7: Corpus-based computational linguistics

Richa and Shahid Mushtaq Bhat
CASE SYNCRETISM IN URDU-HINDI: A CHALLENGE FOR NLP

Imen Ktari
POSTMODIFIERS ACTING AS COMPLEMENTS AND ADJUNCTS IN POPULAR AND ACADEMIC MEDICAL ARTICLES: A GENERATIVE CORPUS-BASED APPROACH

Camino Gutiérrez and Julia Alonso
THE TRACE CORPUS ALIGNER: DEVELOPING A NEW ELECTRONIC TOOL FOR LANGUAGE RESEARCHERS

Biblioteca, 2º piso/floor
Panel 8: Corpus, adquisición y enseñanza de lenguas (Dra. Raquel Criado Sánchez)
Panel 8: Corpora, language acquisition and teaching

Jorge Roselló Verdeguer
EL USO DE LA PUNTUACIÓN EN TEXTOS DE ESTUDIANTES DE EDUCACIÓN SECUNDARIA

Isabel Alonso
LA CONSTRUCCIÓN Y ANÁLISIS DE UN CORPUS DE NARRACIONES DE PROFESORES DE EFL EN PRÁCTICAS: DESCRIPCIÓN, DIFICULTADES Y RETOS

Alazne Ciarra Tejada
ANÁLISIS Y APLICACIÓN DE UN CORPUS CONVERSACIONAL DE ELE PARA EL ESTUDIO Y ENSEÑANZA DE LAS PARTÍCULAS DISCURSIVAS CONVERSACIONALES

Veronica Moreno and Gallardo Beatriz
APLICACIÓN DOCENTE DEL CORPUS PERLA: ENSEÑANZA DEL DÉFICIT LINGÜÍSTICO EN LOGOPEDIA

Viernes/Friday, 8 de abril de 2011
16.00-18.00

Aula multimedia 1, 2º piso/floor
Panel 1: Diseño, elaboración y tipología de corpus (Dr. Francisco Alonso Almeida)
Panel 1: Corpus design, development and typology

Camino Gutiérrez
FROM CATALOGUE TO CORPUS IN DTS: TRANSLATED AND CENSORED CINEMA UNDER FRANCO (TRACECI 1951-1962)

Montserrat Arza Rodríguez
DISEÑO DE UN CORPUS PROSÓDICO ORAL Y REDUCIDO EN EL ÁMBITO DE LA SÍNTESIS DE VOZ

José Manuel Martínez Martínez and Iris Serrat Roozen
RECOPILACIÓN Y TRATAMIENTO SEMIAUTOMATIZADO DE CORPUS PARA EL ESTUDIO DE LA TRADUCCIÓN: PORQUE EL TAMAÑO (Y LA CALIDAD) SÍ QUE IMPORTA

Adonay Custódia Santos Moreira
TURIGAL: COMPILATION OF A PARALLEL CORPUS FOR BILINGUAL TERMINOLOGY EXTRACTION

Atro Voutilainen, Krister Linden and Tanja Purtonen
DESIGNING A DEPENDENCY REPRESENTATION AND GRAMMAR DEFINITION CORPUS FOR FINNISH

Maria Jose Marin Perez and Camino Rea Rizzo
DESIGN AND COMPILATION OF A LEGAL ENGLISH CORPUS BASED ON UK LAW REPORTS: THE PROCESS OF MAKING DECISIONS

Aula de Posgrado, 2º piso/floor

Panel 2: Discurso, análisis literario y corpus (Dr. José Luis Oncins)

Panel 2: Discourse, literary analysis and corpora

Pascual Cantos, Aquilino Sánchez, Raquel Criado and Moisés Almela
COMPUTING READING DIFFICULTY IN ENGLISH LITERATURE (19TH AND 20TH CENTURIES): A CORPUS-BASED STUDY

Leida Maria Monaco
MODALIZING MODERN ENGLISH SCIENTIFIC DISCOURSE: A CORPUS-BASED APPROACH TO MODAL AUXILIARIES IN 18TH -CENTURY LIFE SCIENCES TEXTS (CORUÑA CORPUS)

Anna Ivanova
PRESIDENTIAL SPEECH IN 140 SYMBOLS: A CROSS-CULTURAL ANALYSIS OF TWITTER USE BY BARACK OBAMA & DMITRIY MEDVEDEV

José Santaemilia Ruiz and Sergio Maruenda-Bataller
BUILDING A COMPARABLE CORPUS (ENGLISH-SPANISH) OF NEWSPAPER ARTICLES ON GENDER AND SEXUAL (IN) EQUALITY (GENTEXT): PRESENT AND FUTURE APPLICATIONS IN THE ANALYSIS OF SOCIO-IDEOLOGICAL DISCOURSES

Łukasz Piotr Pakuła
'CIVIL PARTNERSHIP' AND 'GAY MARRIAGE' IN CONTEXT

Carmen Gregori-Signes
COMMUNITY DIGITAL STORIES: A CORPUS ANALYSIS (LA HEMOS QUITADO DE SÁBADO)

Aula multimedia 3, 3er piso/floor

Panel 3: Gramática basada en corpus (Dr. Javier Pérez Guerra)

*Panel 3: Corpus-based grammatical studi*es

Tine Breban, Tom Brzyk, Kristin Davidse and Sigi Vandewinkel
THE FOCUSING USES OF VERY, PURE, SHEER, MERE. A CORPUS-BASED INVESTIGATION OF THEIR FUNCTIONAL-STRUCTURAL STATUS AND THEIR DIACHRONIC DEVELOPMENT

Zixi You
A CORPUS-BASED EXAMINATION OF PERFECTIVE AUXILIARY SELECTION IN OLD JAPANESE

Beatriz Rodríguez Arrizabalaga
ON THE PRODUCTIVITY OF ENGLISH COGNATE OBJECTS. A CORPUS-BASED ANALYSIS

Gonzalo Camiña
NEW NOUNS IN THE SCIENTIFIC REGISTER OF LATE MODERN ENGLISH: A CORPUS-BASED APPROACH

Antonio Vicente Casas Pedrosa
MAIN FEATURES OF ENGLISH PREDICATIVE PREPOSITIONAL PHRASES IN ICE-GB

Salón de Grados, 3er piso/floor

Panel 4: Lexicología y lexicografía basadas en corpus (Dr. Pedro Fuertes Olivera)
Panel 4: Corpus-based lexicology and lexicography

Gema Maiz
THE OLD ENGLISH VERBAL SUFFIX -LÆCAN: DICTIONARY FREQUENCY VS. CORPUS PRODUCTIVITY

Raquel Mateo Mendaza
THE OLD ENGLISH ADJECTIVAL AFFIXES FUL- AND –FUL: A TEXT-BASED ACCOUNT ON PRODUCTIVITY

Carmen Novo Urraca
A TYPOLOGY OF MORPHOLOGICALLY UNRELATED ADJECTIVES IN OLD ENGLISH

Aula Multimedia 2, 2º piso/floor

Panel 5: Corpus, estudios contrastivos y traducción (Dra. Mª Ángeles Gómez)
Panel 5: Corpora, contrastive studies and translation

Marta Fernández-Villanueva Jané and Oliver Strunk
CONECTORES CAUSALES EN LA LENGUA ORAL. UN ANÁLISIS CONTRASTIVO BASADO EN CORPUS ENTRE ALEMÁN Y CATALÁN

Kasper Nijsen
"THIS PAPER ARGUES = DIT ARTIKEL BEWEERT?": IS-AV-CONSTRUCTIONS IN ACADEMIC PROSE TRANSLATION

Laura Cruz-García and Heather Adams
ADDRESSING THE POTENCIAL CUSTOMER IN FINANCIAL ADVERTS: A CONTRASTIVE ANALYSIS IN ENGLISH AND SPANISH

María Cristina Toledo Báez
TRANSLATING RESEARCH ARTICLES FROM SPANISH INTO ENGLISH: A CORPUS-BASED COMPARATIVE ANALYSIS OF THE GENRE

Biblioteca, 2º piso/floor

Panel 8: Corpus, adquisición y enseñanza de lenguas (Dra. Raquel Criado Sánchez)
Panel 8: Corpora, language acquisition and teaching

María José Labrador-Piquer and Pascuala Morote-Magán
LA LENGUA Y LA CULTURA DEL VINO EN LA ENSEÑANZA DE LENGUAS EXTRANJERAS

Ana Valverde-Mateos
USO DE CORPUS ORALES DE APRENDIENTES PARA LA ENSEÑANZA DEL FRANCÉS COMO LENGUA EXTRANJERA

Victoria López
EXPLOTACIÓN DE RECURSOS ON-LINE PARA LA CREACIÓN DE ACTIVIDADES BASADAS EN CORPUS

Carolina Blanes Nadal
LA GESTIÓN DEL CONOCIMIENTO MEDIANTE LAS NUEVAS TECNOLOGÍAS EN LOS CORPORA

Montserrat Mola and Jordi Cicres
PROGRAMACIÓN DIDÁCTICA MEDIANTE EL USO DE CÓRPORA

Cristóbal Lozano and Amaya Mendikoetxea
CEDEL2 (CORPUS ESCRITO DEL ESPAÑOL COMO L2): A LARGE-SCALE CORPUS FOR L2 SPANISH ACQUISITION RESEARCH

Sábado/Saturday, 9 de abril de 2011
10.00-11.30

Biblioteca, 2º piso/floor
Panel 2: Discurso, análisis literario y corpus (Dr. José Luis Oncins)
Panel 2: Discourse, literary analysis and corpora

Milagros del Saz Rubio
AN APPROACH TO NATIVE AND NON-NATIVE WRITERS' USE OF INTERACTIONAL METADISCOURSAL FEATURES IN SCIENTIFIC ABSTRACTS IN ENGLISH WITHIN THE FIELD OF AGRICULTURAL SCIENCES

José Luis Oncins-Martínez
A CORPUS-BASED VIEW OF REPORTING FORMULAE IN DICKENS' NOVELS

Aula multimedia 1, 2º piso/floor
Panel 5: Corpus, estudios contrastivos y traducción (Dra. Mª Ángeles Gómez)
Panel 5: Corpora, contrastive studies and translation

Rosa Currás Móstoles and Miguel Ángel Candel-Mora
MÉTODOS DE LA LINGÜÍSTICA DE CORPUS APLICADOS A LOS ESTUDIOS DESCRIPTIVOS DE TRADUCCIÓN

Cristina Castillo Rodríguez
DETECCIÓN Y CLASIFICACIÓN DE ERRORES DE TRADUCCIÓN DE LAS UNIDADES TERMINOLÓGICAS CONTENIDAS EN UN CORPUS PARALELO MULTILINGÜE DE TURISMO DE SALUD Y BELLEZA

Daniel Gallego-Hernández and Miguel Tolosa-Igualada
ELABORACIÓN DE GLOSARIOS A PARTIR DE CORPUS PARALELOS AD HOC. APLICACIÓN A LA INTERPRETACIÓN DE CONFERENCIAS EN EL ÁMBITO SOCIOECONÓMICO

. Åke Viberg
IMPERSONAL CONSTRUCTIONS IN SWEDISH. A CORPUS-BASED CONTRASTIVE STUDY

Angeles Gómez
CORPUS STUDY BETWEEN THE ENGLISH GERUND AND ITS SPANISH COUNTERPARTS

Iria Gayo and Luz Rello
DIFERENCIAS EN EL PÁRAMETRO PRO-DROP ENTRE PORTUGUÉS BRASILEÑO Y ESPAÑOL UTILIZANDO CORPUS COMPARABLES

Aula de Posgrado, 2º piso/floor

Panel 6: Variación lingüística y corpus (Dra. María José López Couso)

Panel 6: Linguistic variation and corpus

Elisabeth Melguizo Moreno
UNA INVESTIGACIÓN SOCIOLINGÜÍSTICA DE CORPUS EN GRANADA

Maria-Pilar Perea
UN CORPUS DE DIETARIOS DE VIAJES: LOS LÍMITES ENTRE EL DIALECTO Y EL IDIOLECTO

Cristina Illamola
LA INFLUENCIA DE LA L1 EN EL USO DE LA CONSTRUCCIÓN "IR A + INFINITIVO" CON VALOR PROSPECTIVO EN LAS ZONAS BILINGÜES

Pilar Sánchez-García
THE WESTMORELAND DIALECT IN THREE DIALOGUES (1790): THE CONTRIBUTION OF ANN WHEELER'S DIALOGUES TO JOSEPH WRIGHT'S THE ENGLISH DIALECT DICTIONARY

Jordi Cicres
LA LINGÜÍSTICA FORENSE Y EL USO DE LOS CORPUS LINGÜÍSTICOS

Biblioteca, 2º piso/floor

Panel 8: Corpus, adquisición y enseñanza de lenguas (Dra. Raquel Criado Sánchez)

Panel 8: Corpora, language acquisition and teaching

Anna Krasnikova
CORPORA AND TEACHING OF EDITING

Mª Luisa Roca-Varela
CORPORA AS TOOLS AND RESOURCES FOR THE TEACHING OF ENGLISH VOCABULARY

Penny MacDonald, Susana Murcia, Maria Boquera, Ana Botella, Laura Cardona, Rebeca García, Esther Mediero, Michael O'Donnell, Ainhoa Robles and Keith Stuart
ERROR CODING IN THE TREACLE PROJECT

Amaya Mendikoetxea, Cristóbal Lozano and Esther Ferrandis
WHY WE NEED TO COMBINE CORPUS AND EXPERIMENTAL DATA IN L2 ACQUISITION

Anabel Borja Albi, Natividad Juste and Maria Pilar Ordóñez López
EL CORPUS GENTT: LA INTEGRACIÓN DE GÉNERO Y CORPUS EN LA ENSEÑANZA DE LENGUAS PARA FINES ESPECÍFICOS

Aula multimedia 2, 2º piso/floor

Panel 9: Usos y aplicaciones específicas de la lingüística de corpus (Dra. Isabel de la Cruz Cabanillas)

Panel 9: Corpus linguistics: Uses and specific applications

Miguel Lacalle
THE LIMITS BETWEEN AFFIXATION AND COMPOUNDING IN OLD ENGLISH: THE SUFFIX -BORA

Alicia Ricart-Vayá and María Alcantud-Diaz
USING COMPUTER-BASED CORPORA TO CREATE LEARNING MATERIALS FOR TOURISM (ESP)

José María José María Guerrero Triviño, Rafael Martínez Tomás, Mª Carmen Díaz Mardomingo and Herminia Peraita Adrados
MODELO DE RED BAYESIANA BASADO EN UN CORPUS LINGÜÍSTICO DE DEFINICIONES CATEGORIALES APLICADO AL DIAGNÓSTICO DEL DETERIORO SEMÁNTICO COMPATIBLE CON DEMENCIA TIPO ALZHEIMER

Sábado/Saturday, 9 de abril de 2011
12.00-13.00

Biblioteca, 2º piso/floor
Panel 2: Discurso, análisis literario y corpus (Dr. José Luis Oncins)
Panel 2: Discourse, literary analysis and corpora

Gustavo Adolfo Rodríguez Martín
Topic Transition in the Plays of Bernard Shaw: Some Corpus-Based Remarks.

Alcina Sousa and Alda Correia
From Modernity to Post-modernity: conflicting voices in literary discourse - A corpus analysis of you and one

Salón de Grados, 3er piso/floor
Panel 4: Lexicología y lexicografía basadas en corpus (Dr. Pedro Fuertes Olivera)
Panel 4: Corpus-based lexicology and lexicography

Roberto Torre Alonso
THE PREFIX UN- IN THE FORMATION OF OLD ENGLISH NOUNS: COMBINATORIAL PROPERTIES AND CONSTRAINTS

Marta Grochocka
NONCE FORMATIONS AS INDICATORS OF PRODUCTIVE WORD-FORMATION PROCESSES IN ENGLISH

Roberto Therón
ANÁLITICA VISUAL: UN NUEVO ENFOQUE EN LA LINGÜÍSTICA DE CORPUS PARA EL NUEVO DICCIONARIO HISTÓRICO DEL ESPAÑOL

Aula multimedia 1, 2º piso/floor
Panel 6: Variación lingüística y corpus (Dra. María José López Couso)
Panel 6: Language variation and corpus

Iria Romay
A PRELIMINARY STUDY OF NEUTRAL MOTION VERBS IN LOB AND FLOB

Meng Ji
A CORPUS-BASED STUDY OF DIACHRONIC REGISTER VARIATION IN MODERN CHINESE

José Ramón Varela Pérez
NOT-NEGATION AND NO-NEGATION IN CONTEMPORARY SPOKEN BRITISH ENGLISH: A CORPUS-BASED STUDY

Aula multimedia 3, 3er piso/floor

Panel 8: Corpus, adquisición y enseñanza de lenguas (Dra. Raquel Criado Sánchez)
Panel 8: Corpora, language acquisition and teaching

Maria Dolores Garcia-Pastor
LEARNERS' DISAGREEMENTS IN EFL: L2 PRAGMATICS AND THE USE OF A LEARNER CORPUS IN THE LANGUAGE CLASSROOM

Elena Del Olmo Bañuelos, Antonio Moreno Ortiz and María Del Olmo Bañuelos
COMPUTER LEARNER CORPUS (CLC) RESEARCH: UN FUTURO APOYO PARA MATERIALES DIDÁCTICOS BASADOS EN EL MÉTODO CLIL

Mª Isabel Velasco Moreno
INFLUENCIA DEL FEEDBACK EN EL ALUMNADO DE EDUCACIÓN PRIMARIA CON RESPECTO A SU PRODUCCIÓN ORAL EN LENGUA EXTRANJERA

Aula de Posgrado, 2º piso/floor

Panel 9: Usos y aplicaciones específicas de la lingüística de corpus (Dra. Isabel de la Cruz Cabanillas)
Panel 9: Corpus linguistics: Uses and specific applications

Antonio Moreno Ortiz, Chantal Perez Hernandez and Rodrigo Hidalgo Garcia
UTILIZACIÓN DE CORPORA TEXTUALES PARA LA EXTRACCIÓN DE MODIFICADORES CONTEXTUALES DE VALENCIA PARA TAREAS DE ANÁLISIS DE SENTIMIENTO

Katarzyna Marszałek-Kowalewska
CORPUS AND LANGUAGE POLICY: IRANIAN LANGUAGE POLICY TOWARDS ENGLISH LOANWORDS

PÓSTERS/POSTER SESSION

Viernes/Friday 8 de abril. 16.00-18.00

2ª planta/floor Pasillo/Corridor Departamento Lingüística Aplicada

Teresa Marqués Aguado and Laura Esteban Segura
TEXSEN APPLIED TO A CORPUS OF MEDICAL TEXTS IN MIDDLE ENGLISH

David Prieto García-Seco and María Á. López Vallejo
CONFECCIÓN DE UN CORPUS DE FORMACIONES LÉXICAS OCASIONALES PROCEDENTES DE LA LITERATURA DEL SIGLO DE ORO

Elvira Manero Richard
ELABORACIÓN DE UN CORPUS DE TEXTOS PROCEDENTES DE BLOGS PARA EL ESTUDIO DE LA CREACIÓN LÉXICA EN ESPAÑOL

María Á. López Vallejo and David Prieto García-Seco
LA NECESIDAD DE UN CORPUS DOCUMENTAL HETEROGÉNEO EN EL ESTUDIO DE LA TERMINOLOGÍA MILITAR DE LOS SIGLOS XVI Y XVII

Elvira Manero Richard and David Prieto García-Seco
ELABORACIÓN DE UN CORPUS DE UNIDADES FRASEOLÓGICAS A PARTIR DE TEXTOS LITERARIOS

CONFERENCIAS PLENARIAS/PLENARY SESSIONS

JUEVES/THURSDAY 7 DE ABRIL. 10.30-11.30

ETS DE TELECOMUNICACIONES. 3ªPLANTA.
Mike Scott (University of Aston)
Investigating patterns
Although it might seem that the main utility of Corpus Linguistics lies in the power software brings of ploughing through large amount of texts in order to find examples, that is in truth secondary. The main purpose is to identify textual or linguistic patternings. It is through the transformation of a corpus of texts into new forms such as lists, plotted marks, dispersion charts and graphs, word clouds etc. that the linguist perceives patternings that would otherwise be likely to remain invisible. In nature, there are numerous cases of underlying ure, or in our case in large or small amounts of text. The presentation will be illustrated with pattern-identifying techniques available in WordSmith Tools 6.0 (2011).

JUEVES/THURSDAY 7 DE ABRIL. 16.00-17.00

SALON DE GRADOS. DEPARTAMENTO DE LINGÜÍSTICA APLICADA, 3ª PLANTA.
Javier Martín Arista (Universidad de La Rioja)
Uso de corpora lexicográficos y textuales para la elaboración de una base de datos léxica
Esta conferencia versa del uso de The Dictionary of Old English Corpus y los diccionarios de inglés antiguo de Bosworth-Toller, Clark-Hall, Sweet y Toronto con la finalidad de diseñar y elaborar la base de datos léxica del inglés antiguo Nerthus (www.nerthusproject.com). Tras presentar la metodología y análisis empíricos en los que se basa Nerthus, se examina la combinación de un corpus textual con el corpus lexicográfico en tareas como elaborar el listado de entradas, recabar la información pertinente sobre dichas entradas, analizar gradualmente la formación léxica, calcular la productividad de los procesos, identificar préstamos y calcos léxicos, localizar formas reconstruidas, aislar formas no atestiguadas, proporcionar bases de derivación de derivados no tratados como tales por los lexicógrafos, resolver análisis divergentes propuestos por distintas fuentes lexicográficas y establecer los vocalismos y alternancias de los paradigmas derivativos. La conclusión principal que se extrae de la presentación es que la elaboración de una base de datos léxica de una lengua histórica requiere el análisis combinado de corpora lexicográficos y textuales.

VIERNES/FRIDAY 8 DE ABRIL. 12.00-13.00

SALON DE ACTOS ETS TELECOMUNICACIONES, 3ª PLANTA
Susan Hunston (University of Birmingham)
Patterns and Evaluative Meaning
Pattern Grammar offers one way of observing and making explicit the relationship between form and meaning. Alternative similar concepts include Construction Grammar, Local Grammars and Frame Semantics. This paper considers the possibility of exploiting Pattern Grammar to investigate and quantify evaluative meaning,

focusing both on recognised patterns / constructions such as N that or V n as n but also on less frequently considered patterns such as ADJ about n and other adjective patterns. The applications of this approach, such as the automatic recognition of evaluative meaning, will be considered, as will the limitations of the approach.

VIERNES/FRIDAY 8 DE ABRIL. 18.30-19.30

SALON DE GRADOS DEL DEPARTAMENTO DE LINGÜÍSTICA APLICADA. 3ª PLANTA.
Mike O'Donnell (Universidad Autónoma de Madrid)
Using learner corpora to redesign university-level ESL education.
This talk will discuss various means in which a learner corpus collected from ESL students can be used to reshape the educational experience of these students, or those who follow them. Firstly, a learner corpus can provide strong input to the English-teaching curriculum. We can extract 'grammatical profiles' from the learner corpora, showing, for each proficiency level, the grammatical structures which are most critical for the developing students at that level. For this, we can use error annotation, to track what students are doing wrong at each level, and also automatic grammatical analysis, to see what they are getting right. Secondly, an error-annotated learner corpus provides a good basis for material preparation for the teacher. When teaching a particular structure, they can see what kinds of errors the students make, and how frequently. This tells them how much of their teaching materials to dedicate to each problem area, and provides examples to use in those materials. The error corpus can also be used to produce exercises for the students, for instance, asking the students to identify errors, or correct them. Thirdly, we will discuss how the learner profiles mentioned above can be used by an intelligent online exercise system, which offers questions targeted directly at the needs of the student at their current point of language development, and that adapts its conception of the student's proficiency on the basis of the student's responses.

SÁBADO/SATURDAY 9 DE ABRIL. 13.00-14.00

SALON DE GRADOS DEL DEPARTAMENTO DE LINGÜÍSTICA APLICADA. 3ª PLANTA.
Antonio Briz (Universidad de Valencia, Grupo Val.Es.Co.)
Los corpus orales del español: la cualidad y la cantidad de los datos
A partir de una descripción de los corpus orales del español intentaremos mostrar la imagen rica y variada, además de precisa, segura y sistemática que proporciona esta lingüística con corpus. Sin duda, desde el punto de vista metodológico, los corpus son un banco de datos, así como también un banco de pruebas eficaz y natural del lenguaje. Hoy pocos dudan ya de que una hipótesis sin experimentación es meramente especulativa; no llega a ser teoría, si no llega validarse o invalidarse. Y tal validación llega con el trabajo experimental; la investigación se avala con datos reales, depende del corpus y de su observación. Ahora bien, para el avance de esta lingüística y de esta metodología, conviene también debatir sobre algunas cuestiones y problemas que están sin resolver sobre la cantidad o calidad de los datos, las grandes bases de datos o los corpus con objetivos, la suficiencia de los corpus, los accesos a la información, la digitalización, los sistemas de marcación y de transcripción, la explotación, el trabajo de análisis y de abstracción… Más que soluciones, este trabajo intenta plantear

interrogantes sobre lo que se ha hecho y, especialmente, sobre lo que queda por hacer. El siguiente es, sin duda, fundamental: ¿ayudan nuestros análisis a partir de corpus a construir, confirmar, destruir o desconfirmar teorías?

RESÚMENES DE PONENCIAS (ORDEN ALFABÉTICO DEL APELLIDO DEL PRIMER AUTOR)

ABSTRACTS (ALPHABETIC ORDER OF SURNAME)

Alcantud Díaz, María

Panel: 2. Discurso, análisis literario y corpus

VIOLENCE IN CHILDREN'S TALES: A SYSTEMIC CORPUS AND CRITICAL DISCOURSE ANALYSIS OF CINDERELLA

The main aim of this article is to discuss the results achieved after investigating the presence of violence in the brothers Grimm's Cinderella (Tatar 1987,1992,2004); through a corpus-based analysis (Biber 1998) with the intention of finding out what kind of verbal processes predominate in this tale and whether they can be related to violent actions. The tool used for the analysis was WordSmith Tools 5 (Scott, 2010). The study involved first an analysis of frequencies of the lexical units in Cinderella, followed by a comparison of the results obtained in the frequency test to two reference corpora: British national Corpus and Cobuild Concordancer. The analysis was completed with a study of the concordances of some selected words, seeking in detail the context in which they appear. Once the quantitative and qualitative surveys were completed, I then proceeded to analyse the type of verbal processes (Halliday 1994:106-175) extracted from the frequency list. These were classified according to the framework proposed by Downing (2002:111). Thus verbal processes were classified as belonging to six categories: material, mental, verbal, behavioural, existential and relational. After classifying them, these same verbal processes were analysed according to four parameters: who (agent), what (type of action) to whom (affected) and under what circumstances. The results obtained in the frequency and concordance tests of this tale, seemed to indicate that violence is certainly present in Cinderella. The method proved to be a good tool to check whether each character's identity and their social position (power) were somehow related to the infliction of violence. That is, if some characters took the advantage of their predominant position and thus inflicted violence upon other characters. As a general conclusion of the analysis of the results, a tentative proposal could be formulated: that a corpus-based analysis in conjunction with both, a transitivity analysis and a critical discourse analysis, could empirically detect the presence of controversial and polemic topics such as violence in different types of texts. The results could be used as evidence to support a social intervention by means of a linguistic intervention (Graddol and Swann 1989) aimed at decreasing the amount of violent language and situations reproduced in children's tales.

Alcaraz-Mármol,Gema and Lourdes Cerezo-García

Panel: 8. Los córpora y la adquisición y enseñanza del lenguaje

SPECIFIC FREQUENCY AND ITS ROLE IN FOREIGN LANGUAGE VOCABULARY ACQUISITION

Several studies (Saragi et al. 1978; Hulstijn et al. 1996; Reyes 1999; Waring and Takaki 2003; Pigada and Schmitt 2006; Webb 2007) have highlighted the role of specific frequency – i.e, the number of times a word occurs in a text – when it comes to second language vocabulary acquisition. In fact, especially in non-naturalistic contexts of learning, "individual texts within each corpus can vary from one to another and from the overall frequency list which a corpus produces" (Milton 2009: 25). As stated above, the specific frequency of a word may differ from general frequency. Knowing the number of times a word is to be encountered for acquisition would help designers create reading materials adjusted to the learners' needs. Unfortunately, to date, there is no agreement on the number of occurrences that are necessary for acquisition. What is more, we do not even know whether all words need to be encountered the same number of times. A number of studies have focused on this issue (Horst et al. 1998; Laufer 1998; Nation and Wang 1999; Rott 1999). Scholars have tried to determine, as accurately as possible, the number of times a word needs to occur to enable acquisition. What we find in this respect are various different outcomes, ranging from 5 and 20 occurrences. Yet, most of these works are carried out under artificial or laboratory conditions which may be far from mirroring the authentic

learning context. The current study aims to approach the real situation of the classroom. It seeks to define the relationship between specific frequency and vocabulary acquisition within the context of EFL formal instruction. We pursue to answer two research questions:

1) Is there a significant relationship between specific frequency and immediate vocabulary acquisition, regarding receptive and productive knowledge?

2) Is there a significant relationship between specific frequency and mid-term vocabulary retention, regarding receptive and productive knowledge?

In order to achieve our aim, a group of nine-year-old students of EFL in their fourth year of Elementary Education was tested on vocabulary contained in their coursebook. The input for the experiment was taken from Unit 3, which introduced a total of 21 target words (17 nouns and 4 adjectives). These words were classified into three groups, according to their specific frequency. Both written and oral occurrences were taken into consideration. Three weeks before starting Unit 3, target words were pre-tested. Once students had worked with this unit, a receptive and a productive test were distributed both immediately after finishing the unit, and three months later. Results show that the effect of specific frequency on vocabulary learning differs depending on the moment this learning is assessed, that is, whether it is tested just immediately after dealing with vocabulary or some months later.

Alcina Sousa and Alda Correia

Panel: 2. Discurso, análisis literario y corpus

FROM MODERNITY TO POST-MODERNITY: CONFLICTING VOICES IN LITERARY DISCOURSE - A CORPUS ANALYSIS OF YOU AND ONE

From modernity to postmodern discourse, places, landscapes and people are aesthetically perceived and reshaped, within the perspective of alterity/otherness, upon which one constructs the image of "one's own" and the "other" in a dialogical game of mirrors. This paper discusses the possibilities of a corpus analysis applied to literary interpretation. It is, thus, our goal to present our preliminary findings of a work in progress intended to disambiguate some pronominal references, i.e. one / you, as they occur in prose fiction, namely in two of Virginia Woolf's and Hugo Hamilton's novels. These involve readers in a dialogic interpretation of the text's "polyglossia", either conveying the generic pronoun reference or the protagonist's inner voice. In Hamilton's The Speckled People (2003), the shifting pronominal reference I/you points to a multitude of pulls either inwards or outwards be it in the sphere of the individual and the community to which he belongs, or in the physical space. Very often in the novel, the focaliser / protagonist presents an alternate view to mainstream ideology, reinforced by the generic pronoun reference you. By contrast, one occurs more frequently in Virginia Woolf's texts. This evidences a linguistic/stylistic choice conforming to patterns of use from modernity to post-modernity which draw the attention to her way of conceiving her feminist project and a postmodern aesthetics.This analysis will benefit from a multi-layered interpretive framework drawing on discourse analysis, and corpus-based approaches, particularly in that it unpacks ways in which writers make use of linguistic structures. The analysis of the collocational meaning (in Alan Partington 1998: 9-10) "can provide powerful support for a reader's intuition". Consequently, the reader is challenged "to explore new kinds of identity and forms of relationship" or, according to Martin Montgomery et al. (1995: 121), "to see the world from unfamiliar and revealing angles… by subverting the commonsense bonds between utterances and their situations of use".

Almela, Ángela and Gema Alcaraz

Panel: 2. Discurso, análisis literario y corpus

MEASURING WILDE'S STYLE: AN APPLICATION OF COMPUTER STYLOMETRY TO A LITERARY CORPUS

The study of authorial style in literary and non-literary works has always been a staple in humanities. It is generally assumed by researchers in the field that people have a characteristic pattern of language use that can be detected in their way of speaking and in their writings, and the first applications of this theory aimed at authorship attribution. As Juola puts it, "[d]isputes about the ownership of words have been around for as long as words themselves could be owned" (2008: 237). In the era of personal computers and corpus linguistics, the study of style in language has seen its greatest development, giving rise to the discipline known as "Computer Stylometry". Within this field, simple statistics have been combined successfully, being the most notable example of this the Delta method (Burrows, 2002). This method is considered to produce very positive results (Cantos et al., 2010); hence the fact that authors such as Argamon (2008) and Hoover (2004, 2004a) have proposed interesting modifications of the method. This method has been commonly evaluated on literary texts, such as English poems and novels, by different authors. More recently, it has also been used to discover patterns of similarity and difference in works by the same author, in order to detect stylistic variation throughout their work and to examine how patterns in dialogue are used to individualize characters, that is to say, to construct their idiolect. Even though this kind of computational testing provides a sound basis for an emerging discipline, there are so far just two studies which explore characters' idiolects, and none of them include Delta procedure in their research methodology. First of all, Rybicky (2006) studied character idiolects in Henryk Sienkiewicz's trilogy and their two English translations. Subsequently, Rybicki (2008) has conducted an examination of the idiolects of the characters of Shakespeare's Hamlet, in which nine randomly-selected translations into various languages are compared by means of Multidimensional Scaling graphs of characters' speech, based on the relative frequencies of the most common words. In view of the preceding discussion, this work is intended as a contribution to the available empirical knowledge on the computational stylometric analysis of literature through the application of Delta method. Specifically, we will delve into characterisation in Oscar Wilde's oeuvre, since, to the best of our knowledge, this celebrated writer has not been object of any computational stylistic analysis yet. For the discrimination of characters within the same play, we have performed Delta and Delta Prime analysis of the idiolects in English originals. Specifically, the spreadsheet have listed the 100 most common words in descending order of their frequency in the corresponding subset, shown their mean frequencies as percentages of that set, presented the corresponding standard deviations, and given z-scores representing their divergences from the means of the other subsets. In addition, a Wilcoxon signed-rank test has been performed. The results do suggest idiolectal divergences among several characters and certain linguistic patterns shared by characters of the same social group.

Almela, Ángela and Samuel Gracia

Panel: 5. Corpus, estudios contrastivos y traducción

EL GUIÓN CINEMATOGRÁFICO COMO CORPUS: UN ESTUDIO CONTRASTIVO ENTRE EL ESPAÑOL CASTIZO DE ALMODÓVAR Y SU TRADUCCIÓN AL INGLÉS

En España, país doblador por excelencia, se ha prestado especial atención a la traducción audiovisual enfocada al doblaje de productos importados, mientras que la traducción a otra lengua de productos audiovisuales españoles ha recibido menor atención desde el punto de vista de la investigación lingüística (Chaume, 2005). Tal es precisamente la motivación del presente estudio, que presenta un análisis contrastivo de la traducción al inglés de películas realizadas por uno de los cineastas españoles más internacionales: Pedro Almodóvar. Diversos elementos lingüísticos y traductológicos de su obra han sido ya objeto de estudio, como en el trabajo de investigación realizado por Baldi (2004). De mayor relevancia para nuestro estudio es el trabajo llevado a cabo por Moreno (2006), en el que se emplea un corpus paralelo de cinco películas originales de Almodóvar y su traducción al inglés en forma de subtítulos, siendo la más reciente la que constituye el objeto de estudio de la presente investigación: La Mala Educación. A partir de dicho corpus hemos realizado el análisis contrastivo de ambas versiones, original y traducida, desde el aporte teórico de traductólogos como Hernández Sacristán (1996). Para tal fin, se ha empleado un método de estudio híbrido. En un primer estadio, se ha llevado a cabo un análisis cuantitativo por medio de ciertas herramientas de la lingüística de corpus que ofrecen información relevante sobre el texto analizado, tales como el estudio de las palabras clave. Este estudio de tipo

cuantitativo ha sido interpretado y matizado por medio de un análisis posterior de naturaleza cualitativa del corpus meta, en el que se han observado los elementos culturales dentro del contexto histórico-cultural de la época franquista y la Transición y la manera en que éstos se han trasvasado a la lengua meta. Además de ello, nos hemos centrado en la subtitulación al inglés del slang de dicho film, y más especialmente en las expresiones expletivas dentro de un registro coloquial, ya que creemos firmemente que debe prestarse mayor atención desde un punto de vista académico al tratamiento de las palabras tabú, puesto que éstas, al formar parte de la vida diaria, se retratan en los productos audiovisuales con los que un traductor debe trabajar (Rojo y Valenzuela, 2000). Los resultados preliminares muestran que, en lo que respecta al uso tan característico que de su identidad sexual hacen los personajes, el traductor ha tratado de subsanar la ausencia de flexión propia del inglés con algún mecanismo de compensación. Además, los apelativos tabú tan frecuentemente empleados en el original y tan característicos del habla de los personajes principales no siempre encuentran en la versión subtitulada el mejor equivalente. Conviene destacar que no todos los casos en los que la traducción no se corresponde con el original se justifican por la restricción espacial de los subtítulos, lo que apunta a la necesidad de que el traductor tenga muy presentes la dimensión cultural y la dimensión pragmática para plasmar en la lengua meta el mensaje original con la misma frescura.

Almela, Moisés

Panel: 4. Lexicología y lexicografía basadas en córpora

FROM COLLOCATION TO INTER-COLLOCATION: DEVELOPING A DYNAMIC APPROACH TO COMBINATORIAL LEXICOGRAPHY

The lexicographical treatment of collocation has been focused on descriptions of dependencies between words. This involves typically the combination of a node and its collocates. This perspective of analysis can be described as “intra-collocational”, because it is centered on the analysis of internal relationships within a bigram. There are, however, strong reasons to argue that the intra-collocational perspective in combinatorial lexicography is incomplete and sometimes even misleading. Recent studies in corpus-based lexicology have suggested that the collocational profile of a node is in part shaped by interdependencies among its collocates (Cantos & Sánchez, 2001; Sánchez et al., 2007; Almela et al., 2011). Therefore, in order to increase the accuracy of collocational descriptions, the intra-collocational perspective should be complemented with an “inter-collocational” analysis that is, with an analysis of the way in which different collocations of a word exert an influence on each other. The existence of an interaction between two or more collocations is observed wherever the association strength of a node-collocate pair is reinforced or weakened as a result of the effect exerted by other neighboring elements. Thus, given a node word W and three of its collocates (C1, C2, C3), the probability of finding C1 in the context of W can be increased or decreased by the presence C2 or C3. To put it more formally, we can say that the intra-collocational perspective is concerned with dependencies of the following form: W|C1, W|C2, C1|W, C2|W, etc., while the inter-collocational perspective is concerned with dependencies of a more complex form, namely: (W,C1)|C2, (W,C1)|C3, (W,C2)|C3, etc. For example, the likelihood that the noun policy functions as a direct object of the verb review is higher when it is modified by adjectives such as existing or current in comparison with cases in which policy is modified by local; and conversely, the probability of finding other verbal collocates, such as implement and develop, in the context of policy is higher when the adjective is local in comparison with situations in which the adjective is existing or current. Thus, we can say that existing and current are “co-collocates” of the pair review + policy, but not of the pair implement + policy. This paper submits a proposal for introducing inter-collocational information into electronic collocation dictionaries. There are, of course, serious objections to the incorporation of this type of contextual data in printed dictionaries, due to obvious limitations of space. However, in electronic lexicography these practical difficulties can be resolved with the help of expanded menus and user interfaces. The central idea of this paper is that by creating a more dynamic design of lexical entries in electronic combinatorial dictionaries it is possible to include more detailed contextual information, especially inter-collocational relations. The advantages over more conventional approaches to combinatorial lexicography will be illustrated with reference to lexical entries for the nouns policy and control.

Alonso Campo, Araceli

Panel: 4. Lexicología y lexicografía basadas en córpora

COLLOCATIONAL NETWORKS Y EL USO 'ESPECIALIZADO' Y 'GENERAL' DE LAS UNIDADES LÉXICAS: EL CASO DE AQUALEXIC

Las palabras se usan en todo tipo de situaciones situaciones marcadas por la temática y situaciones no marcadas y palabras que se utilizan en un ámbito temático específico pasan a formar parte del vocabulario general del hablante y viceversa. Todas estas palabras acaban formando parte del acervo lingüístico del hablante medio. Como indica Lara (1990), el diccionario ha de recopilar no sólo el vocabulario general, sino también aquel vocabulario más especializado que ha pasado, desde la experiencia social, a formar parte del idiolecto de un hablante. Todo este trasvase entre unidades hace a veces difícil la delimitación –si es que la hay- entre lo general y lo especializado. De hecho, coincidimos con otros autores (Meyer 2000; Hunston y Sinclair 2003; Ahumada 2004; ten Hacken 2008; Williams y Millon 2010, entre otros) en que no se puede establecer realmente una dicotomía entre lo "general" y lo "especializado", sino que se ha de tratar en términos de un continuum. Uno de los ámbitos donde precisamente los límites son difusos, por ser un ámbito multidimensional, de gran difusión e interés social es el del medio ambiente. De hecho, se puede observar una falta de concreción en la representación de las unidades relativas a este campo de conocimiento en los diccionarios (Alonso 2008, 2009; Alonso y DeCesaris 2007; Marimón 2008), por lo que son necesarios estudios teóricos y descriptivos que permitan desarrollar y aplicar una metodología para determinar los diferentes grados de especificidad que presentan estas unidades según los diferentes contexto de uso y poder así caracterizar el léxico del medio ambiente y mejorar su representación lexicográfica. El trabajo que presentamos tiene su origen en el estudio realizado en el marco de la tesis doctoral (Alonso 2009) y forma parte de un proyecto de investigación en curso sobre la caracterización del léxico del medio ambiente mediante la aplicación de la Theory of Norms and Exploitations y Corpus Pattern Analysis (Hanks 2004 y en prensa) y el uso de collocational networks y collocational resonance (Williams 1998, 2002, 2006, 2008a, 2008b; Williams y Millon 2010). Este estudio se centra, concretamente, en el uso de collocational networks y en mostrar cómo estas redes de colocaciones facilitan la observación de los usos "generales" y "especializados" de las unidades léxicas relativas al medio ambiente, así como de las relaciones sintagmáticas y paradigmáticas que se establecen entre las diferentes unidades, lo cual permite determinar nuevas pautas de representación de estas unidades léxicas en el diccionario.

Alonso,Isabel

Panel: 8. Los córpora y la adquisición y enseñanza del lenguaje

LA CONSTRUCCIÓN Y ANÁLISIS DE UN CORPUS DE NARRACIONES DE PROFESORES DE EFL EN PRÁCTICAS: DESCRIPCIÓN, DIFICULTADES Y RETOS.

Esta comunicación describe los avances realizados en el proyecto de construcción de un corpus de narraciones escritas de profesores de EFL en prácticas en la Universidad Autónoma de Madrid, así como las dificultades surgidas durante una primera anotación de los recursos discursivos y léxico-gramaticales utilizados para la expresión de juicios y valoraciones en relación a la profesión de la enseñanza. El corpus se nutre principalmente de los diarios de clase y las reflexiones escritas que los alumnos de la Facultad de Formación de Profesorado y Educación de la UAM redactan durante su periodo de prácticas en los centros públicos de la Comunidad de Madrid. El fin último de este proyecto es la elaboración de una teoría discursivo-funcional (Halliday, 1985/1994; Martin y Rose, 2003/2007) sobre el perfil identitario profesional de los futuros profesores de EFL en Primaria y Secundaria y sobre cómo éste evoluciona a través de las diferentes fases de las prácticas.

Alonso-almeida, Francisco and Ivalla Ortega-Barrera

Panel: 5. Corpus, estudios contrastivos y traducción

EVIDENTIALITY AND EPISTEMIC MODALITY IN ENGLISH AND SPANISH LEGAL SCIENTIFIC DISCOURSE: A CORPUS-BASED STUDY

This paper explores the concepts of evidentiality and epistemic modality in a corpus of English and Spanish legal scientific discourse. The data for analysis is taken from Evycorpe, a database of English scientific papers in the fields of computing, medicine and law published between 1998-2008. For the present work, we only focus on the legal part of the corpus, but the results will be implemented with the other two register subdomains in the future. The Spanish legal corpus has been gathered for this contrastive study following the same Evycorpe criteria of compilation. The notions of epistemic modality and evidentiality are differently treated in the literature (Dendale and Tasmowski 2001). Whereas for some scholars evidentiality represents a subdomain of epistemic modality (Chafe 1986, Palmer 2001), there are others who consider evidentiality as an independent category (Cornillie 2009). Epistemic modality is strongly connected to the idea of "truth" and the authors' responsibility concerning their statements (Traugott 1989; Sweetser 1990; Stukker, Sanders and Verhagen 2009). Evidentiality is seen as the coding of the authors' "source of knowledge", and this may eventually imply differing degrees of certainty concerning the proposition manifested (Carretero 2004). In this paper, we follow an intersective approach and, although both categories are kept theoretically distinct, they undergo functional overlapping. The use of these strategies might be indexical of the authors' position and intention in discourse (Marín Arrese 2009). This said, our main objectives are (1) to identify and classify epistemic and evidential markers in the corpus, and (2) to describe their frequency of occurrence in each language subcorpus and their functions mainly as stance markers. The paper concludes that epistemic markers appear in higher frequency in the English texts, whereas the Spanish ones tend to show more examples of evidential strategies, although in both cases these marker types aim to be manifestation of face-saving expressions (Brown & Levinson 1978), among other pragmatic effects.

Alvarez Mosquera, Pedro

Panel: 9. Usos específicos de la Lingüística de Corpus

TESTING THE EXCEPTION: AN ANALYSIS OF EMINEM'S LANGUAGE USES FROM A CORPUS-BASED APPROACH.

Eminem's presence in the hip-hop scene has been controversial ever since he burst into the music world in the late 90's (Bozza 2003: 93). His exceptional success as a Caucasian in a predominantly African American genre is reflected in the number of records he sold and the significant support he garnered from influential figures in the hip-hop world. While Eminem was attacked by those who accused him of being a product of the music industry for the purpose of selling millions of records to the white market, others defended him for his genuine talent as a rapper. Analyzing rap's linguistic component, which plays a central role in the genre, is a way to potentially evaluate Eminem's authenticity as a rapper in an objective manner. By maintaining a sociolinguistic approach, we used Wordsmith Tools to process Eminem's language choices in his album, The Marshall Mathers LP, launched in 2000, and we compared them with contemporary African American rapper Jay Dilla's album, Welcome 2 Detroit, released in 2001. Analyzing similarly sized corpora from two rappers who share the same relative age, city of origin, and gender, allows us to focus on ethnicity and language as the center of this study. Our results emphasize significant similarities in how both rappers use rap as a communicative device, following specific linguistic patterns ascribed to the role and function of the African griot in the African American tradition. However, important differences were also noted. The limited references to the central concept of community, and the absolute absence of the term nigger in Eminem's corpus (among other features), set him apart from the African American group and put him closer in line with the corpus associated with other Caucasian rappers (Álvarez-Mosquera 2010). Finally, our data also illustrates that authenticity is a highly disputed quality in rap music. Rap is intrinsically interwoven with ethno-cultural patterns as a result of the Black Experience (Rose 1994: 123), which has made African American rappers'

linguistic uses culturally coded and specific, undermining Caucasian rappers' attempts to sound more authentic.

Arza Rodríguez, Montserrat

Panel: 1. Diseño, compilación y tipos de córpora

DISEÑO DE UN CORPUS PROSÓDICO ORAL Y REDUCIDO EN EL ÁMBITO DE LA SÍNTESIS DE VOZ

Las propuestas metodológicas dirigidas a la elaboración de un corpus intentan ajustarse, en cada caso, a los objetivos del trabajo de investigación que se esté llevando a cabo. En este estudio se tratarán con cierto detenimiento las orientaciones metodológicas que se siguen en los trabajos dedicados a la entonación. Lo cierto es que son más abundantes las reflexiones sobre la creación de corpus generales dedicados al estudio de la entonación, que las reflexiones que versan sobre corpus específicos destinados al estudio de dominios prosódicos. Si es cierto que abundan las reflexiones sobre la confección de corpus destinados al estudio de dominios prosódicos, también es cierto que al emprender la elaboración de éstos se cometen errores. Un corpus entonativo implica una gran asistematicidad en cuanto a la impredicibilidad de los resultados, ya que dependen de la voz de cada informante. Así, la primera disyuntiva es decidirse por un corpus de habla espontánea o un corpus creado ad hoc. El corpus hablado, aunque aportaría mayor naturalidad, es incontrolable en cuanto a número de sílabas, acentos y estructuras sintácticas, entre otros. Por este motivo, es bastante frecuente que en los estudios de entonación aplicados a la síntesis de voz, los corpus sean creados artificialmente y leídos. Al tiempo que se decide el tipo de corpus ya se va confeccionando su contenido y cómo se va a obtener. Se debe decidir si contendrá frases aisladas o texto discursivo y cuál será la estructura interna sintáctica y prosódica de cada uno de ellos. Nuestro propósito es llevar a cabo el diseño de un corpus para el estudio de la entonación, de forma que se pueda extraer información que permita hacer una sistematización de fenómenos prosódicos y sintácticos. El fin que se persigue con la creación de reglas más o menos fijas, es conseguir una mayor naturalidad de la voz artificial, lo que supone que el corpus ha de tener unas especificidades técnicas de partida, que serán explicadas en el apartado correspondiente. Así pues, se presenta una metodología de diseño de corpus prosódico reducido, tratando el tipo de informantes, el modo de grabación, el tamaño del corpus, el tipo de estructuras prosódicas y sintácticas elegidas y todos aquellos puntos necesarios para conseguir nuestro propósito.

Aurrekoetxea, Gotzon

Panel: 7. Lingüística computacional basada en corpus

"CORPUSLEM" UNA HERRAMIENTA PARA LA CONVERSIÓN DE CORPUS TEXTUALES EN DATOS

Los corpus textuales presentan nuevas oportunidades para el estudio de la lengua. Hasta hace pocos años el estudio de los corpus textuales presentaba dificultades para el análisis automatizado. Estas dificultades van aminorándose con la aparición de nuevas técnicas y nuevas herramientas de codificación de los mismos (TEI Speech). Los corpus orales (Spoken corpus) que recogen textos de variedad estándar son objeto de análisis en todos los idiomas desarrollados. Los corpus orales de habla espontánea de variedades dialectales presentan más dificultades para su análisis automatizado, por la carencia de herramientas adecuadas para su explotación. El grupo de investigación "Eudia" de la Universidad del País Vasco/Euskal Herriko Unibertsitatea ha creado una herramienta de conversión de corpus textuales a base de datos, denominado "CorpusLem". Es una herramienta online que no necesita instalación local y que se puede acceder desde cualquier lugar con conexión a Internet. La versión actual está diseñada para distintas lenguas (vasco, inglés, español, francés, catalán...), tanto en el interfaz como en los contenidos. Esta herramienta, por una parte, convierte documentos de distintos formados (.doc, .odt o .txt) en datos estructurados en formato MySQL; por otra parte, proporciona un índice alfabético de todas las palabras, agrupando las palabras por semejanza o correspondencia ortográfica, y propone un lema para todas las variantes encontradas, con la opción de modificarlo. Con objeto de una

correcta corrección de los lemas y determinar su contenido semántico, la herramienta proporciona el contexto de cada palabra. El usuario tiene la opción de realizar las correcciones tanto en la misma herramienta como en su propio ordenador, con la opción de descargar. Y una vez corregido implementarlo de nuevo. La herramienta está diseñada para albergar diferentes proyectos y soporta más de un usuario por cada proyecto, pudiendo acceder cada uno de ellos a más de un proyecto, todos ellos autorizados por el gestor de la herramienta. El programa puede actuar con textos en variedad estándar o variedades dialectales, en grafía actualizada o grafía original de los textos, en cuyo caso han de ser aplicadas una serie de reglas creadas por el usuario.

Ávila Martín, Carmen and Ramón Martí Solano

Panel: 4. Lexicología y lexicografía basadas en córpora

EL ANÁLISIS DISCURSIVO DE LA VIOLENCIA A TRAVÉS DE UN CORPUS ESPECÍFICO

El análisis del discurso mediático tiene en la actualidad herramientas que nos permiten la realización de análisis empíricos cuantitativamente más documentados que los análisis tradicionales. El análisis léxico de las coocurrencias discursivas nos aporta datos de interés para interpretar cómo se construyen los discursos. Para ello utilizaremos la creación de corpus específicos que nos aportan datos objetivos sobre la utilización discursiva de las unidades analizadas. El presente trabajo se enmarca en el proyecto de investigación ALEC denominado "Relaciones de género y prácticas sociales: red Iberoamericana /Europa/Caribe ALEC" de la Universidad de Limoges (Francia). El trabajo de investigación se propone analizar el tratamiento mediático de la violencia en la prensa británica y española a través del estudio de los elementos léxicos que las expresan. La alianza de la lingüística del corpus y del análisis textual permitirá obtener suficientes datos empíricos que servirán para explicar estos fenómenos y para situarlos en sus contextos interlingüísticos y extralingüísticos. Para ello realizaremos el análisis de algunas lexías recurrentes en la prensa británica y española. En el caso del inglés la lexía hate crime hace referencia a los delitos motivados por la hostilidad hacia la víctima como miembro de un grupo social. En la prensa española la utilización de expresiones de la violencia contra las minorías se expresa a través de la unidad léxica acoso. Para su estudio, crearemos un corpus específico de 100 000 palabras y, después del análisis cuantitativo, haremos un análisis cualitativo de los contextos léxicos más frecuentes de estas unidades léxicas. Esta primera etapa de la investigación será ampliada con un estudio comparativo del tratamiento mediático de diversas formas de violencia en la prensa española y en la prensa francesa. Al análisis cuantitativo de los datos lingüísticos procedentes de los corpus de prensa de los tres países implicados le seguirá un análisis cualitativo y comparativo de los resultados obtenidos. El objetivo de esta investigación es mostrar cuáles son los contextos y las asociaciones lingüísticas relacionados con la violencia infligida principalmente a las mujeres, a los adolescentes, a los minusválidos, a los homosexuales y a las minorías raciales y religiosas.

Bartholamei Junior, Lautenai Antonio

Panel: 1. Diseño, compilación y tipos de córpora

PEPCO: DESIGNING A PARALLEL AND COMPARABLE TRANSLATIONAL CORPUS IN BRAZIL

Brazilian studies in translation have been growing in last years, as well the use of corpus tools to help researchers. The used of tools provided by corpus linguist are often used to help translators in their researches or training. PEPCo (pepco.ufsc.br) was designed to be a tool which can help scholars and researchers in their task of create e explore texts in the corpus. Design process of PEPCo was carried out in two steps: (i) corpus design, i.e., text selection, representativeness; and (ii) development of tools, i.e., the use of a MySQL database and PHP scripting language, designing of an interface for querying and retrieving data from the corpus using HTML, CSS and JavaScript. Most used tools provided by PEPCo are parallel concordances, monolingual concordances, word-lists, n-grams, and PEPCo Builder. PEPCo Builder is a tool that makes easier the corpus compilation by the user. The user does not need to have

technical knowledge on corpus tools and scripting, he/she only needs a pre-aligned parallel text in a text processor and all sentences/paragraphs need to match in source and target texts. Then, both source and target text are uploaded using a web form and user receive an unique corpus ID by an e-mail provided in the form and then can access his/her own corpus through a web page. The result (in progress) is a parallel corpus of about 3 million words and a comparable corpus of about 5 million words which could be useful for many researchers in translation studies in Brazil. Most researches using PEPCo are related to translation studies and translational phenomena emerging from a compiled corpus. Popular genres in PEPCo are Fantasy, Science-Fiction, Medical and Academic Texts. Corpus tools provide filters to user search for specific texts, genres, period, authors, translators, publishers. Also, users can specify to query only on source text, target text or both. In case of querying for both texts, user can define a node for source text and another one for target text. PEPCo is used by students and teachers to researches and translator's training in Southern Brazil. PEPCo developers and users are always integrating new resources provided to aid each new research.

Bengoechea, Mercedes and José Simón

Panel: 6. Corpus y variación lingüística

FEMINIST LANGUAGE REFORM IN SPANISH ADVERTISING. A CORPUS-BASED RESEARCH

Within the framework of a broader research project, we have examined the evolution of gender adscape along the past years. Our aim was to investigate to what extent non-sexist language has been used in the advertisements published during October 2007 in the most influential newspaper in Spain, El País, which is also the one with the widest readership. We have collected two samples in three years: the first one corresponds to October 2007 and the second to October 2010. In addition, all advertising received in a middle-class home in Madrid during the same period was equally collected and analysed. A key element in our survey was the corpus we created with our samples. In order to streamline the study, a database was created in which, once scanned, some 700 ads were registered using a double format, jpg images and pdf, together with the text of the advert. Among common data (date, section, page, etc.), we also registered the type of product or service advertised. Then, in the same database we annotated them according to gender treatment in verbal usage. In this paper we present the results of the first phase of our study, which corresponds to the advertising in el País during the month of October 2007, with particular emphasis on the corpus methodology we have followed.

Blanes Nadal, Carolina

Panel: 8. Los córpora y la adquisición y enseñanza del lenguaje

LA GESTION DEL CONOCIMIENTO MEDIANTE LAS NUEVAS TECNOLOGIAS EN LOS CORPORA.

El conocimiento representa uno de los valores más importantes para lograr el éxito sostenible en cualquier organización. La habilidad para adquirir información, transformarla en conocimiento e incorporarlo en las unidades productivas, constituye un pilar vital para poder enfrentarse a la sociedad, preservar su posición y alcanzar un estado de mejora continuado. Pero para proceder a una correcta implantación de un sistema de gestión del conocimiento como código de conducta profesional, hace falta en primer lugar indicar cuales van a ser las herramientas metodológicas, para posteriormente clasificarlas en básicas y avanzadas. ¿Pero cómo podemos plasmar la necesidad de utilizar la gestión del conocimiento mediante las nuevas tecnologías en los córpora? Para entender esto debemos hablar de los córpora electrónicos basados en las nuevas tecnologías. El recuento estadístico de las unidades léxicas aparecidas en los corpus lingüísticos de las lenguas extranjeras da lugar a diccionarios o listados de frecuencias léxicas, conocidos también como vocabularios básicos. Mediante el recuento de las unidades léxicas se pretende dar cuenta del vocabulario más empleado por las personas que utilizan una lengua. Así pues con esta participación intentamos apuntar justamente a la necesidad de establecer la

enseñanza sistemática del vocabulario basado en el criterio de la frecuencia teniendo en cuenta la gestión del conocimiento.

Borja Albi,Anabel, Natividad Juste and Maria Pilar Ordóñez López

Panel: 8. Los córpora y la adquisición y enseñanza del lenguaje

EL CORPUS GENTT: LA INTEGRACIÓN DE GÉNERO Y CORPUS EN LA ENSEÑANZA DE LENGUAS PARA FINES ESPECÍFICOS

La aplicación del concepto de género a la enseñanza de lenguas y, especialmente a la enseñanza de lenguas para fines específicos, se ha convertido desde la década de los ochenta en una de las líneas de investigación más dinámicas en la investigación sobre géneros. Bazerman (1988, 2000), Bhatia (1993, 2004) y Swales (1990, 2004), entre otros, destacan la importancia de comprender los códigos comunicativos específicos de la cultura de las distintas áreas de especialización así como la estructura de los géneros característicos de dichas áreas. La investigación llevada a cabo por el grupo GENTT (Géneros Textuales para la Traducción) se centra en el estudio multilingüe de los géneros en contextos profesionales especializados, en el ámbito jurídico, médico y técnico, ámbitos que ocupan una posición clave en la enseñanza de la lengua para fines específicos. El corpus GENTT, resultado de la labor de compilación realizada durante la última década, es un corpus multilingüe (catalán, castellano, inglés, alemán y francés) de géneros especializados, de los tres ámbitos profesionales mencionados anteriormente. La utilización del corpus GENTT en la enseñanza de lenguas para fines específicos nos permite poner al alcance del alumnado modelos y patrones textuales que le proporcionan referencias textuales, conceptuales, lingüísticas y terminológicas. A su vez, el corpus GENTT, construido en base al concepto de género, proporciona información formal, comunicativa y cognitiva de los géneros que contiene. Así, pretendemos que el corpus se convierta en un sistema de gestión del conocimiento especializado a través del género, con directa aplicación tanto para la docencia como para los profesionales que trabajan con géneros especializados (Borja, 2005). El corpus GENTT constituye un entorno de trabajo colaborativo que permite a los distintos tipos de usuarios alimentar, buscar y gestionar el corpus online de manera autónoma, lo que lo convierte en una herramienta efectiva –dinámica e interactiva— de enseñanza-aprendizaje. Con este trabajo pretendemos poner de manifiesto cómo la incorporación del corpus GENTT así como del enfoque basado en el género a la enseñanza de lenguas para fines específicos, en este caso el inglés económico-jurídico, nos ayuda a superar las críticas dirigidas hacia el uso de metodologías bottom-up en la lingüística de corpus y, por otro lado, hacia el uso descontextualizado de los datos contenidos en el corpus. Este trabajo incluye una serie de actividades prácticas, basadas en el uso del corpus GENTT, con las que ilustraremos la utilización del corpus en el aula de inglés económico-jurídico.

Borosi, Bernadette

Panel: 4. Lexicología y lexicografía basadas en córpora

CORPUS PARALELOS ALINEADOS: SEGMENTACIÓN TEXTUAL CON FINES LEXICOGRÁFICOS

Las nuevas tecnologías en las últimas décadas han originado la transformación metodológica tanto en la presentación como en la elaboración de los productos lexicográficos. Nuevas subdisciplinas de la lingüística aplicada, como es, por ejemplo, la lingüística de corpus, con la añadida posibilidad técnica de gestionar y analizar un sinfín de datos desde múltiples puntos de vista, se convierten en fuentes y herramientas imprescindibles de la lexicografía moderna, potenciando el carácter multidisciplinar de la misma. En nuestra comunicación, trabajando con la combinación de la lengua española y húngara, presentamos las ideas fundamentales de una propuesta de metodología para la segmentación textual en corpus bilingües alineados y el registro de las unidades delimitadas en una base de datos bilingüe con fines lexicográficos. A partir de los textos paralelos bilingües en línea que nos brinda la legislación europea en la temática de medio ambiente, elaboramos nuestro corpus paralelo alineado,

facilitándonos esta adaptación de formato la delimitación sistemática de cadenas textuales en uno de los idiomas y sus candidatos a equivalentes en el otro. Considerando que la principal función de un diccionario bilingüe es tender un puente entre los dos idiomas salvando las diferencias lingüísticas en todos los niveles, el estudio contrastivo de las estructuras lingüísticas subyacentes, en relación con las necesidades comunicativas de los posibles usuarios, nos llevará a delimitar las mono- y polilexías que pasarán a formar parte de la base de datos bilingüe que alimentará el diccionario. Mediante una pequeña demostración se ejemplifican las relaciones semántico-funcionales entre los distintos tipos de segmentos, visualizando las unidades léxicas de forma contrastiva en los dos idiomas, que nos inducirá a reflexionar sobre las posibles estructuras de presentación lexicográfica de las mismas. Si bien nuestra comunicación se centra en el estudio de las diferencias lingüísticas que se puedan dar entre la lengua española y húngara, y en las posibles soluciones para su registro lexicográfico, entendemos que el método propuesto puede servir como procedimiento de compilación y análisis comparativo, aplicable para distintas combinaciones de idiomas.

Bouda, Peter

Panel: 3. Estudios gramaticales basados en córpora

LANGUAGE DOCUMENTATION CORPORA IN DESCRIPTIVE LINGUISTICS

The role of corpora in the creation of descriptive grammars has gained a lot of attention in the last decades. Still, only few grammars directly refer to corpus analysis as a main mean to extract the linguistic information they present. In recent years the usage of software tools in language documentation projects generated a new source of linguistic data, that will be used to compile decriptive grammars for lesser-used and endangered language in the future. It is the goal of this paper to present a software solution to search and analyze annotated corpora that were created in language documentation projects. The software is especially designed for the application with DOBES corpora, but may be extended to other kinds of corpora later on. In the first part, I will outline some of the questions a descriptive linguist will pose to a corpus when he is in the process of writing a grammar. Those questions resulted in a typology of searches the linguist needs to apply to a corpus, in order to extract the information about grammatical types and relations on all linguistic levels. This typology was the basis to create a list of requirements for a software tool that is currently used in two language documentation projects. Real-world examples from those projects will be presented to show how to derive grammatical descriptions from corpora through search and analysis within the software tool. In the second part I would like to present the technical solution in detail, a preliminary version of a database/concordancing software specifically designed to fulfil the functions and principles outlined in the first part. It supports the Elan and Toolbox file format, two of the main software packages used in DOBES documentation projects. Those data files typically contain transcriptions, morpho-syntactic annotations and translations, which are accessible through a search interface within the software. Search results are displayed with full interlinear data, so that context and annotation data are displayed to the user. The software implements the search strategies that were derived from the requirements outlined in the first part, for example successive searches on previous search results, or search for classes of words, morphemes, glosses, etc. extracted from fieldwork sketches. Parts of the corpus or search results may directly be published in hypertext documents, i.e. in digital grammars, by a simple copy and paste procedure. Later versions of the software will allow publishing whole corpora in a standardized XML format based on the Corpus Encoding Standard with fixed URLs that allow access and links to the data on a simple web server. Depending on access restrictions the underlying data files may also be accessed directly from the DOBES archive at the Max-Planck-Institute in Nijmegen or other archives.

Breban, Tine, Tom Brzyk, Kristin Davidse and Sigi Vandewinkel

Panel: 3. Estudios gramaticales basados en córpora

THE FOCUSING USES OF VERY, PURE, SHEER, MERE. A CORPUS-BASED INVESTIGATION OF THEIR FUNCTIONAL-STRUCTURAL STATUS AND THEIR DIACHRONIC DEVELOPMENT.

The starting point of this paper is formed by the problems posed by a little described element of the English NP, viz. the prenominal focusing adjective. It occurs in postdeterminer position and its semantics are similar to focusing adverbs, such as inclusive 'even' (1, 2) and exclusive 'only' (3), manifesting wide (1,2) and narrow scope (3).

(1) Many commentators feel that the deadly cocktail of drugs, guns and Aids sweeping inner city America is threatening the very existence of Afro-Americans.

(2) Anyone who freezes with fright at the mere sight of the dentist's chair will be pleased to know that you can now tune into something more relaxing than a screeching drill.

(3) We had been hoping for it to coincide with Keats's birthday, but you can imagine how hard it proved to cram 12 whole quatrains into a mere four hours.

The central question is whether they are best treated as secondary determiners (Bolinger 1968, Adamson 2000) because of their structural position and general 'reference-modifying' function, or as a type of emphasizer (Quirk et al 1985, Vandewinkel & Davidse 2008) because of their inherent or latent scalarity. We will approach this issue from a diachronic angle, studying the focusing uses from their earliest appearances on (in which they may still be entwined with secondary determiner and/or degree modifier uses) and analysing the diachronic changes they underwent to clarify their status in contemporary English. This investigation will be based on systematic qualitative and quantitative analysis of historical and contemporary corpusdata with the adjectives very, pure, sheer and mere. Extractions were made from the Helsinki corpus (750-1150), the Penn-Helsinki Parsed Corpora of Middle English (1150-1500) and Early Modern English (1500-1710), the Corpus of Late Modern English Texts (1710-1920), and the COBUILD corpus (1993-). The first diachronic question that we want to settle is whether the focusing uses of these adjectives emerged as a subtype of the degree modifier use or of the secondary determiner use. We will answer this question by charting the relative proportions of these three uses throughout the main periods of English and by investigating the bridging contexts (Wilkins & Evans 2000) in which one reading is a focusing reading. Our second diachronic question pertains to the pragmatic-semantic development of the various focusing uses: exclusive, inclusive, particularizing; wide vs. narrow scope; scalar vs. non-scalar (König 1989, Nevalainen 1991 1994, Eckardt forthc.). Despite the original association of pure, sheer and mere with exclusive meaning and of very with inclusive meaning, they all developed focusing uses unpredicted by their lexical meaning. Based on close analysis of all the relevant contextualized examples, we will trace paths of change, based both on the more general meaning shifts established in pragmatic theory and on the gradual extension of collocates of the adjectives in their focusing use. Our data-based reconstruction of these collocational histories will allow us to assess the importance in "emergent grammar" of collocational persistence and extension, with the language community's awareness of "prior text" as an important source of grammaticalization (Hopper 1998). This extensive qualitative and quantitative study of corpus data will allow us to develop an historically-informed description of the neglected prenominal focuser function of adjectives. We will situate the focuser function in relation to subjective and intersubjective meaning and scalarity in the whole English NP.

Brett, David and Antonio Pinna

Panel: 9. Usos específicos de la Lingüística de Corpus

LEXICAL BUNDLES IN US PRESIDENTIAL SPEECHES: A CORPUS-DRIVEN STUDY OF B. CLINTON'S, G.W. BUSH'S AND B. OBAMA'S ADDRESSES

In this paper we investigate patterns of variability in lexical bundles in a corpus of US presidential addresses and compare our findings with those reported in the literature concerning other fields of discourse. In our study we adopted Biber's (2009) methodological approach which he used to

investigate variability within multi-word units using two corpora: a 4.5-million-word corpus of American English conversation; and a 5.3-million-word corpus of academic prose. Initially, the corpora were searched for 4-grams, discarding sequences with a frequency of less than 10 occurrences per million words. Each corpus was then searched for a series of sequences composed of three of the components of each 4-gram, allowing variability in the fourth slot, e.g. *234, 1*34 etc. If the token in a given slot in each 4-gram composed less than 50% of the results for that slot, the slot was deemed to be variable, as opposed to fixed, and marked with an asterisk. This procedure permitted the identification of typical patterns of variability in the formulaic sequences across the two corpora. For example, internal variability in one slot (1*34/12*4) was seen to be relatively common in Academic Prose, whereas initial and final variability (*23*) was more frequent in the conversation data. The corpus which we have used for this study is composed of US presidential addresses and remarks delivered by B. Clinton (1993-2000), G.W. Bush (2001-2008) and B. Obama (2009-2010). As a macro-genre Presidential speeches are monologic texts characterized by being usually prepared to be recited in public. They could therefore be expected to contain features of both written and oral language, possibly tending towards the oral end of the cline. This led us to speculate that our data would fit this picture by showing patterns of variability which positioned Presidential speeches as more or less evenly straddling the oral-written divide as defined by Biber's (2009) findings.

Broadly speaking, the presidential data patterns display greater similarity to those of conversation, rather than academic writing: internal variation (12*4/1*34 and 1*3*/*2*4), which is characteristic of academic writing, is infrequent in both; conversely, variation in the external slots (123*/*234) is common in both (particularly so in the former), while being considerably less frequent in academic prose. However, a marked difference may be noted in the proportions of wholly invariable patterns (1234). In Biber's conversation and academic prose data, these represent merely 7% and 8.5% of the total patterns, respectively. On the other hand, this pattern constitutes no less than c. 21% of the total in our presidential data. Further analysis reveals considerable variation among presidents: Bush's use of such patterns is remarkably high in comparison to his immediate predecessor and successor. On the whole, we may conclude by observing that the presidential address data displays far higher levels of formulaicity than the reference genres, as almost 55% of the patterns are of three types: 1234, 123* and *234.

Brown, David and Laura Aull

Panel: 2. Discurso, análisis literario y corpus

"TOUGH GUYS" AND "CATFIGHT CRAZY": A CORPUS-BASED ANALYSIS OF GENDER REPRESENTATIONS IN SPORTS REPORTAGE

This study uses a corpus-based approach to investigate the discursive representations of athletes and their connection to ideologies of gender. To carry out this investigation, we have compiled two specialized corpora: one containing press accounts covering a fight that took place between the Detroit Shock and the Los Angeles Sparks of the Women's National Basketball Association (WNBA) and the other containing press accounts covering a fight between the Detroit Pistons and the Indiana Pacers of the National Basketball Association (NBA). In our analysis, we find that the narratives in the NBA corpus are constructed around the allocation of blame, often focusing on the role of a particular player, Ron Artest, and the behavior of fans. In contrast, the narratives in the WNBA corpus are often constructed around the fight's effect on the league—in particular whether the fight will bring positive or negative attention. In addition, the WNBA corpus contains a large number of gender-marked tokens (e.g., female, men, girls, boys, daughters, femininity) indicating that the reportage often generalizes the specifics of the WNBA fight to construct broader representations of gender and gender norms. The results of the study are facilitated by the analysis of keywords, token frequencies, and collocations, as well as comparisons of linguistic features of our corpora to sports reportage features more generally evidenced in the Corpus of Contemporary American English. The purpose of our investigation is two-fold. First we want to interrogate the intersections of gender, sport, and language, in order to illustrate how sport can be a productive site for exploring issues related to language and ideology, but also that it is importantly

implicated in social constructions of gender. Second, we want to contribute to the growing body of research using corpora both large (e.g., Rayson, Leech, and Hodges 1997; Schmid and Fauth 2003) and specialized (e.g., Motschenbacher 2009) to show, in Baker's (2008: 74) words, "the untapped potential" of corpus linguistics in the study of language and gender.

Calzada Perez, Maria

Panel: 5. Corpus, estudios contrastivos y traducción

ANÁLISIS CRÍTICOS DE DISCURSOS PARLAMENTARIOS EUROPEOS. DESDE LA TEXTURA AL CONTEXTO CON ECPC

CPC (European Comparable and Parallel Corpora)* es un archivo de corpus (comparables y paralelos)compuestos por discursos procedentes del Parlamento Europeo (PE), el Congreso de los Diputados (CD) y la Cámara de los Comunes británica (HC). El archivo, compilado desde la Universitat Jaume I (Castellón de la Plana, España) por el grupo homónimo (cuya coordinación recae en María Calzada Pérez y entre cuyas filas cabe destacar a investigadores de la talla de Mona Baker, Dorothy Kenny y Silvia Bernardini), se inspira en importantes proyectos europeos como OPUS (Open Source Parallel Corpus, Tiedemann 2009), TEC (Translational English Corpus, Laviosa 1998, Baker 1999) y ENPC (English Norwegian Parallel Corpus, Johansson 1997, 2007). Sin embargo, su metodología de compilación y etiquetado es, en gran medida, innovadora. Mediante la automatización de tareas de etiquetado en XML, los corpus registran parámetros textuales y metatextuales (género de los oradores, afiliación política, referencias generacionales, función desempeñada, lengua de expresión original, etc.) que permiten la exploración "inteligente" de fenómenos lingüísticos y traductores vinculados con discursos parlamentarios. Y es precisamente esta metodología compiladora la que posibilita procesos de análisis crítico que enlazan textura con contexto (adentrándose en el proceloso campo de la visión, la cosmovisión y la ideología) a través del examen de las diversas prácticas discursivas parlamentarias. Tras una breve descripción de la naturaleza y método compilador de ECPC (apartado 1), el presente trabajo se propone desarrollar análisis contrastivos con el material del archivo ECPC (apartado 2) que, partiendo de listados de frecuencia, palabras clave y examen de concordancias al más puro estilo sinclairiano, estudien comportamientos de oradores parlamentarios de diversas clases (hombres frente a mujeres, conservadores frente a progresistas, cargos gubernamentales concretos y oradores específicos) y los cotejen (cuando parezca oportuno) con las traducciones de las intervenciones originales. Para ello, se hará uso de las premisas investigadoras de estudiosos de corpus como Sinclair (2003), Xiao y McEnery (2006), y Scott y Tribble (2006); de traductólogos como Tognini-Bonelli 2001; y de defensores de los CADS (Computer-Assisted Discourse Studies) como Bayley (2004) y Partington et al. (2004), entre otros. Tras estos estudios críticos, la presente comunicación demuestra que es posible y fructífero replicar los modos de compilación y análisis desarrollados con ECPC. Así, se describe, brevemente y a modo ilustrativo, el corpus monolingüe en inglés OBAHILL (apartado 3), consecuencia inmediata del trabajo con ECPC, que recoge los discursos que emitieron Barack Obama y Hilary Clinton durante las pasadas elecciones primarias de EEUU. Con ánimo ejemplificador, nunca exhaustivo, se repasan algunos de los resultados que se han obtenido tras su análisis para enfatizar la posibilidad de réplica de las fases investigadoras de ECPC. El artículo propone una serie de conclusiones (apartado 4) que reflexionan acerca de las metodología de compilación y análisis de los corpus y de las posibilidades que estos ofrecen para retratar la sociedad actual.

Camiña, Gonzalo

Panel: 3. Estudios gramaticales basados en córpora

NEW NOUNS IN THE SCIENTIFIC REGISTER OF LATE MODERN ENGLISH: A CORPUS-BASED APPROACH.

This paper revises word-formation processes in the scientific register of English in the eighteenth century. Using corpus-based methodology, the parser Coruña Corpus Tool and other data processing software, it aims at providing relative frequency patterns to illustrate the most productive processes to

coin new nouns in the fields of astronomy and philosophy in the Late Modern English period. To achieve this we have analysed over 400,000 lexical items corresponding to two sub-corpora contained in the Coruña Corpus of English Scientific Writing, i.e. the Corpus of English Texts on Astronomy (CETA), and the Corpus of English Philosophical Texts (CEPhiT). By means of quantifiable data, we intend to measure the productivity of the different units and processes involved in the coining of nouns. Besides, we will offer two different approaches to the linguistic material in the corpus: on the one hand, diachronic evaluations of the entire corpus that may define the features of the scientific register in general; on the other hand, a synchronic comparison of the two disciplines that may identify unique morphological characteristics inherent to each of them.

Candel-Mora, Miguel Angel and Chelo Vargas Sierra

Panel: 5. Corpus, estudios contrastivos y traducción

ANÁLISIS DE LA PRODUCCIÓN INVESTIGADORA EN LINGÜÍSTICA DE CORPUS APLICADA A LA TRADUCCIÓN

En un momento en el que la Lingüística de Corpus aparece consolidada como disciplina de investigación en lingüística y cuando ha extendido la mayoría de sus métodos y técnicas de análisis y estudio del comportamiento del lenguaje a otras disciplinas como la lexicología, la enseñanza de lenguas y la traducción, junto con los continuos avances en proceso de datos, capacidad de almacenaje y disponibilidad de cada vez más datos en formato electrónico, parece el momento propicio para una llevar a cabo una reflexión sobre la producción científica y las líneas de investigación de la lingüística de Corpus con una de esas disciplinas: la Traducción. Este trabajo propone un estudio bibliográfico de la literatura en traducción durante los últimos 5 años con el fin de identificar las aportaciones de la lingüística de corpus a la investigación en traducción, y sus aplicaciones. A partir de la información registrada en dos bases de datos bibliográficas BITRA y Translation Studies Abstracts Online de St. Jerome Publishing, se seleccionan las publicaciones en las que confluyen ambas disciplinas y se analizan diferentes variables con el fin de extraer, entre otras cosas, las líneas de investigación, los pares de lenguas, las líneas aplicadas y las teóricas, y en definitiva la adaptación de los métodos de la lingüística de Corpus a la investigación en traducción. Los resultados ponen de manifiesto el auge y la consolidación de los métodos de la lingüística de corpus en la investigación en traducción y perfilan con precisión la evolución de esa relación multidisciplinar, incluso se observa la asimilación de una terminología propia que se ha adaptado de la Lingüística de Corpus aplicada a la Traducción.

Cantos, Pascual, Aquilino Sánchez, Raquel Criado and Moisés Almela

Panel: 2. Discurso, análisis literario y corpus

COMPUTING READING DIFFICULTY IN ENGLISH LITERATURE (19TH AND 20TH CENTURIES): A CORPUS-BASED STUDY

Readability indices (Coleman & Liau, 1975) have been widely used in order to measure textual difficulty. They have proven to be consistent and reliable (Smith & Kincaid, 1970) and can be truly useful for the automatic classification of texts, especially within the language teaching discipline. Among other applications, they allow for the previous determination of the difficulty level of texts without even the need of reading them through. The Automated Readability Index (ARI, hereafter) was originally used to produce an approximate representation of the US grade level needed to comprehend a specific text. Its calculation is based on two ratios: word length (in characters) and sentence length (in words). In this research we shall enlarge its domain and apply the ARI, one of the most used readability indices, to English prose. The aim of this investigation is threefold: first, examining and determining the degree of reading difficulty, ARI, of the 19th and 20th century novels specified below; second, by means of the data obtained, trying to classify and arrange them according to their degrees of reading difficulty, both

individually and chronologically; and third, correlating the data with the English language proficiency level of Spanish university students of Grado de Estudios Ingleses (compliant with the European Space for Higher Education, active from the academic year 2009-2010) and the Licenciatura de Filología Inglesa (the old Curricula Plan, to become extinct in 2012-2013). Methodologically, we shall calculate the ARI indices of the text corpus consisting of 17 novels by renowned British writers in the 19th and 20th centuries. The authors and novels selected are: (a) from the 19th century, Charles Dickens (Oliver Twist, David Copperfield, A Tale of Two Cities, Great Expectations, Our Mutual Friend); Emily Brontë (Wuthering Heights); Charlotte Brontë (Jane Eyre); George Eliot (Middlemarch); William Makepeace Thackeray (Vanity Fair), and Thomas Hardy (Far from the Madding Crowd); (b) from the 20th century, Joseph Conrad (Heart of Darkness); David Herbert Richards Lawrence (Sons and Lovers); Virginia Wolf (To the Lighthouse); Aldous Huxley (Brave New World); Graham Greene (The Heart of the Matter); George Orwell (1984) and William Golding (Lord of the Flies). Next, we shall arrange the resulting data in a hierarchical way, by means of a cluster analysis, in order to establish the similarities/divergences encountered among the authors/novels/centuries. Finally, we shall correlate the data with the proficiency level of English of our Spanish university students of Grado de Estudios Ingleses and Licenciatura de Filología Inglesa. We are confident that the ARI indices, the clustering of the authors/novels and the resulting correlation might highlight in some way whether the proficiency level of English of our students is up to the degree of difficulty of the English novels recommended in the curricula at our universities. The practical results can be taken as a reference for deciding on the ordering and grading of the literary texts studied along the degree of Grado de Estudios Ingleses.

Carmo, Felix

Panel: 9. Usos específicos de la Lingüística de Corpus

WHAT DO COMPRESSION ALGORITHMS TELL US ABOUT LANGUAGE?

In recent years, there have been many studies in the domain of machine learning regarding the application of compression algorithms to detecting patterns in text and languages. These studies have shown that using these algorithms on unsupervised experiments with different models of data compression can identify regularities which often elude a linguistic analysis. We will present some of these studies, such as the one by Cilibrasi and Vitanyi (2004), in which this method was used in conjunction with clustering techniques to discriminate and group languages by language family, literary works by author, and literary translations by translator. However, these studies pose a lot of questions on what enables a technology which clearly has no linguistic knowledge, such as data compression, to identify distinguishing features in complex computer objects like natural language texts. Mahoney (2010) claims that text compression is a hard Artificial Intelligence problem, due to the difficulty in reaching an adequate language model, and then coding it efficiently. Some of the questions we pose relate to the capacity of these algorithms to distinguish between a string of characters and a meaningfully organised phrase of words. We also question which mathematical parameters improve an algorithm's efficiency in detecting text regularities. Ultimately, these questions try to understand what these algorithms show us about language. We will include some of our own research with a parallel corpus, which shows that, even in small-scale research, compression algorithms are efficient tools for finding textual relations that we would not expect from a mathematical analysis tool. In our experiment, compression algorithms highlight fundamental differences between English and Portuguese translations. There is however, a lot of work to be done in order to identify which text features lead to the algorithm detecting these differences. This is an ongoing project, and a few new stages of work may be added to the presentation.

Carrió Pastor, Maria Luisa and Eva Mestre Mestre

Panel: 9. Usos específicos de la Lingüística de Corpus

THE USE OF CORPUS ANALYSIS TO MANAGE FOREIGN LANGUAGE ACQUISITION IN A BILINGUAL COMMUNITY

Worldwide communication is possible nowadays using English as an international language or lingua franca. English is used in countries with different cultural backgrounds, a fact which affects in the use of pragmatic strategies. On occasions, authors who communicate in a foreign language cannot avoid the use of structures that are more common in their mother tongue (L1). In a monolingual community, language errors could be caused by L1 interference; nevertheless the methodology applied in error analysis and in corpus compilation could vary in a bilingual community. The linguistic status of three languages in contact may not be equal; consequently ideological, linguistic and social factors could influence language acquisition. The main objective of this paper was to find out if the general methodology used for corpora classification is adequate for a corpus of learners with different linguistic background. Furthermore, we analysed if the increasing importance of English as a lingua franca influences students to consider local or national languages less important when developing professional skills. In this article, we used corpus analysis methodology to determine if learners whose mother tongues were Spanish and Catalan varied their errors when learning English. Foreign language acquisition is a universal concept although we consider that the proficiency of some skills could depend on the mother tongue of the learner. In order to analyse the corpora, which included the errors of English texts written by students whose mother tongue was Catalan or Spanish, we conducted an experimental research that included the categories of communicative, grammatical and lexical errors. The results showed that students with different cultural backgrounds produced a dissimilar amount of communicative and lexical errors while both groups produced a similar amount of grammatical errors. As a consequence of this research, we concluded that the methodology used to detect errors should vary depending on the linguistic background of learners.

Casas Pedrosa, Antonio Vicente

Panel: 3. Estudios gramaticales basados en córpora

MAIN FEATURES OF ENGLISH PREDICATIVE PREPOSITIONAL PHRASES IN ICE-GB

This paper is aimed at identifying which are the main characteristics of those English prepositional phrases which perform the function of subject complement in the British component of ICE. Such is the case of “She first fell in love with Will when she was eighteen, and she adores him still” (ICE-GB:W2F-019 #47:1). After introducing the notions of prepositional phrase and subject complement, these structures will be described from the morphological, syntactic, semantic, lexical, and socio-pragmatic points of view and examples will be provided. Although in terms of frequency this is not the syntactic function prepositional phrases more often perform, they are taken into account because of their complexity and due to the lack of detailed analyses. In most cases they are described as isolated examples and this phenomenon is not considered to be a very productive one. Morphologically speaking, prepositional phrases can be defined as those phrases headed by a preposition which requires another unit following it and acting as its complement. Even though there is a wide range of units that can perform the function of complement of a preposition, attention will only be paid to noun phrases. They can be very simple (consisting of a single noun, as “on fire”) or more complex (for instance, “in the pink of health”). From the syntactic point of view, prepositional phrases usually perform the functions of adverbial, postmodifier of noun phrases and complement of adjective and prepositional phrases. Nevertheless, they can also behave as subject and object complements: “That is of no importance” (Quirk et alii, 1985: 732) and “I don’t consider myself at risk” (op.cit.: 733). As far as semantics is concerned, when acting as subject and object complements, prepositional phrases convey meanings which are similar to those of adjectives, since they express qualities or characteristics. Thus “on cloud nine” and “in the doldrums” can be replaced by “very happy” and “depressed”, respectively. Lexically speaking, some of the examples under analysis are idiomatic, their meaning being metaphorical. Such is the case of “(be) on tenterhooks”, which is defined as follows in OALD6 (1340) as “(to be) very anxious or excited while you are waiting to find out sth or see what will happen”. More information is provided as regards its origin: “From tenterhook, a hook which in the past was used to keep material stretched on a drying frame

during manufacture". As far as socio-pragmatics is concerned, sometimes these structures are selected because they allow speakers to express the same meaning by means of a lower number of words. This is the case of "in hand", defined as "receiving attention and being dealt with" (OALD5: 537). Moreover many of these structures are labelled as "colloquial", "informal", "old-fashioned", or "slang" in dictionaries. In some cases they can even convey two different meanings, one being neutral and the other, informal; the phrase "on the job" in OALD6 (697), is thus defined as "while doing a particular job" and "(BrE, slang) Having sex".

Castellón, Irene, German Rigau, Salvador Climent, Marta Coll-Florit and Marina Lloberes

Panel: 7. Lingüística computacional basada en corpus

ANOTACIÓN SEMÁNTICA DEL CORPUS SENSEM

Este trabajo presenta la anotación semántica de los núcleos argumentales de SENSEM (Vázquez y Fernández 2008): sus objetivos, metodología, proceso, criterios y resultados. SENSEM es un banco de datos compuesto por un corpus del español y una base de datos interrelacionados. En su estado previo el corpus estaba etiquetado a nivel sintáctico en su totalidad, y a nivel semántico por lo concerniente a la semántica del núcleo verbal (Alonso et al. 2007). En esta investigación se ha afrontado la anotación semántica de los argumentos, centrándose en la de sus núcleos nominales, con el objetivo final de adquirir las preferencias semánticas de los predicados verbales. Las categorías lexico-semánticas utilizadas para la anotación son las de WordNet 1.6 del español (WNe) (Vossen ed. 1998), habiéndose usado asimismo como base de conocimiento de apoyo el Multilingual Central Repository (Atserias et al. 2004) el cual integra WNe con múltiples ontologías de propósito general. La anotación ha sido realizada por un equipo de 6 lingüistas y ha proporcionado los siguientes resultados:

- La anotación de 23.307 formas correspondientes a 3.693 lemas (82,6% del volumen total del corpus).

- Un conjunto de criterios de anotación, incluyendo instrucciones para anotadores, procedimiento de anotación de nombres propios, soluciones a problemas habituales y, especialmente, criterios para la desambiguación de significados.

- Un análisis en profundidad de la adecuación de WNe para la anotación semántica

- Un conjunto de propuestas para la solución de los problemas derivados de inadecuación de WNe: agrupación de sentidos y operadores especiales de anotación.

La principal característica de SENSEM es su diseño especialmente orientado a la estructura sintáctico-semántica del verbo, lo que se concreta en una constitución representativa y equilibrada de lemas y ocurrencias verbales y una anotación manual, detallada y en profundidad de las unidades verbales. La metodología de anotación utilizada incorpora la experiencia de Agirre et. al (2006) en la creación del corpus anotado del euskera Eusemcor. Se dividió en una fase de preparación técnica —preparación y anotación morfosintáctica del corpus mediante FreeLing (Padró et. al 2010) y adaptación de la interfaz de Eusemcor— y una fase de anotación en forma de secuencia de ciclos de etiquetado y establecimiento de criterios y de acuerdo entre anotadores y árbitros. Esta fase ha implicado el análisis de la adecuación de WNe para la anotación semántica de nombres, profundizándose en el ya iniciado por el grupo en Carrera et al. (2008). Como resultado, esta investigación ha generado instrucciones generales de anotación (e.g. aspectos de estuctura léxico-semántica a considerar, fuentes primordiales de consulta), criterios de anotación y soluciones a problemas más frecuentes (e.g. aplicación de categorías MUC a la anotación de nombres propios, tratamiento de significados metafóricos o metonímicos, anotación de unidades multipalabra, de variantes morfológicas...). De forma especial se han definido criterios para la desambiguación de significados de WNe, sin duda el problema fundamental del proceso. El corpus SENSEM está a libre disposición de la comunidad bajo licencia GPL.

Castillo Rodríguez, Cristina

Panel: 5. Corpus, estudios contrastivos y traducción

DETECCIÓN Y CLASIFICACIÓN DE ERRORES DE TRADUCCIÓN DE LAS UNIDADES TERMINOLÓGICAS CONTENIDAS EN UN CORPUS PARALELO MULTILINGÜE DE TURISMO DE SALUD Y BELLEZA

Desde el conocido "boom turístico", que se produjo durante el periodo comprendido entre los años 50 y 70 (Vogeler y Hernández, 2000), el turismo se ha convertido en una fuerza económica y una realidad social poderosa que ha suscitado el interés de los estados en tanto instrumento para alcanzar objetivos culturales, sociales, educativos e, incluso, políticos. Analizar el impacto económico del turismo es analizar el lugar que ocupa el turismo en el comercio internacional, así como en las economías nacionales. Pero principalmente el turismo es una industria, puesto que se trata de un conjunto de actividades que tienen por objeto la explotación de las riquezas turísticas, así como la transformación de los recursos humanos, de capital y de materias primas, tanto en servicios como en productos. Es en este punto donde la práctica de la traducción se hace eco, entendiéndose ésta como actividad vital para poder trasladar y comunicar hacia otras lenguas todas esas riquezas turísticas, fundamentales para producir negocio. Lamentablemente, las traducciones turísticas realizadas en nuestro país adolecen de no transmitir toda esta realidad, dando lugar a traducciones de muy mala calidad, que a veces, incluso, no llegan a transmitir ni la mitad del mensaje turístico en sí al potencial turista internacional. Además, con la llegada de Internet todas estas traducciones han quedado expuestas a la vista de todos, con lo cual continuamente corren el riesgo de ser evaluadas por cualquiera de los potenciales usuarios, dejando a España en una posición inversamente proporcional al lugar que ocupa como destino vacacional preferido mundialmente (OMT, 2009). El objetivo de esta investigación es analizar la calidad de las traducciones publicadas en la red del material promocional del segmento del turismo de salud y belleza. Para ello se ha compilado un corpus paralelo multilingüe integrado por textos originales (TO) escritos en lengua española y sus textos traducidos o textos meta (TM) al inglés, francés e italiano sobre el segmento del turismo objeto de estudio —para la recopilación de los textos se seguirá la metodología protocolizada por Seghiri Domínguez (2006)—. Sin embargo, previamente al análisis contrastivo y a la evaluación de la calidad de las traducciones se llevará a cabo la tarea de alineación de estos textos traducidos con su original (cf. Castillo Rodríguez, 2010) para, posteriormente, gestionarlos con el programa de gestión de corpus paralelo ParaConc. Una vez que se integren los TO y los TM en este programa de gestión de corpus, se analizarán los textos en aras de clasificar los errores de traducción más frecuentes cometidos en lo que respecta a las unidades terminológicas traducidas en los pares de lenguas español inglés, español francés y español italiano.

Cheng, Su-han and Jeng-yih Hsu

Panel: 8. Los córpora y la adquisición y enseñanza del lenguaje

A CORPUS-BASED STUDY OF THE VOCABULARY USE IN AN ENGLISH NEWSPAPER

In an attempt to create a journalistic English word list (JEWL), this study examines the most frequently occurring words in a 20 million-word journalistic English corpus (JEC) collected from an English newspaper published in Taiwan between 2002 and 2009. Adopting a commercial concordance software package, ConcGram 1.0, this study is able to report its findings on the statistically frequent words, collocations, and four-word lexical bundles. Altogether, 411 word families, which accounts for 4.66 % of total running words in the entire journalistic English corpus, 100 most frequent collocations of the 7 types (i.e., verb-noun, adjective-noun, noun-verb, noun 1of noun 2, adverb-adjective, verb-adverb, and noun-noun), and 100 most frequent four-word lexical bundles are recorded in this study. This journalistic English word list (JEWL), containing perhaps the most important single-word items, the top 100 collocations, and the most commonly seen four-word bundles, may serve as a guide not only for instructors in designing textbooks and courses for journalistic English but also for learners in setting

their goals for vocabulary learning and improving their understanding and comprehension of media English.

Ciarra Tejada, Alazne

Panel: 8. Los córpora y la adquisición y enseñanza del lenguaje

ANÁLISIS Y APLICACIÓN DE UN CORPUS CONVERSACIONAL DE ELE PARA EL ESTUDIO Y ENSEÑANZA DE LAS PARTÍCULAS DISCURSIVAS CONVERSACIONALES

Resumen: En el presente trabajo, en primer lugar, se define y describe el corpus así como su proceso de elaboración. En segundo lugar, se comenta el análisis que se ha llevado a cabo sobre el mismo. El interés se ha centrado en el análisis de los marcadores discursivos conversacionales, y concretamente en la partícula claro. Así, se observa y recoge la aparición de esta partícula en conversaciones de alumnos extranjeros de nivel B1-B2 y se analizan su posición en el enunciado, sus distintos valores y su frecuencia de uso. Finalmente, se comparan estos resultados con el recuento del uso de la misma partícula en hablantes nativos de español (corpus Val.Es.Co.). El estudio de la frecuencia de uso de los marcadores conversacionales en hablantes nativos y en alumnos de español permitirá clasificar el grupo de partículas discursivas conversacionales en niveles de manera que pueda priorizarse el aprendizaje de unas sobre otras según el nivel de dominio lingüístico del estudiante, desde B1 hasta C2. Esta propuesta contribuye al análisis del discurso conversacional oral en segundas lenguas así como en particular del español como LE.

Cicres, Jordi

Panel: 6. Corpus y variación lingüística

LA LINGÜÍSTICA FORENSE Y EL USO DE LOS CORPUS LINGÜÍSTICOS

En este artículo se discute acerca del uso de las distintas clases de corpus lingüísticos en la lingüística forense, tanto desde el punto de vista de la investigación como el de la práctica profesional. Típicamente, el trabajo del lingüista forense consiste en la comparación técnica de textos (orales o escritos). Por un lado, dispone de un corpus de textos dubitados (cuya autoría se desconoce) y un corpus de textos indubitados (cuya autoría es conocida). Los textos dubitados son aquellos sobre los que el perito lingüista forense debe dictaminar. Es impescindible, pues, disponer de textos indubitados que se correspondan al posible autor o a los posibles autores de los primeros. Sin embargo, el perito debe de utilizar también corpus de referencia que le permitan decidir acerca de la rareza o idiosincrasia de las variables presentes en los corpus dubitado e indubitado. La definición de estos corpus de referencia es altamente compleja (y no siempre es posible, tanto por las dificultades técnicas como por la disponibilidad de tiempo). Sin embargo, estos corpus permiten calcular, para algunos parámetros, ratios de verisimilitud (likelihood ratios) dentro del marco bayesiano, con lo que el perito dispone de información muy valiosa que le permite llegar a conclusiones más fiables en sus dictámenes. En este artículo se presentan ejemplos del uso de los distintos corpora en lingüística forense (tanto en casos de determinación o atribución de autoría de textos orales y escritos, como para el análisis del plagio) y se discuten las dificultades metodológicas relacionadas con los distintos tipos de corpora en lingüística forense.

Conejero, Marta, Asunción Jaime and Debra Westall

Panel: 1. Diseño, compilación y tipos de córpora

NIP & TUCK: A CORPUS-BASED QUALITATIVE TYPOLOGY FOR CONCISION IN SCIENTIFIC WRITING

Among the challenges facing researchers who use English as an Academic Language (EAL) is finding out how to publish in the high-impact journals edited by the predominately English-language industry. For many EAL researchers in Spain, the problem is compounded by the discourse community's standards, especially since different fields and different journals seem to have different standards regarding 'linguistically-acceptable' manuscripts. Recently, the terms of acceptability have become much more demanding as editors expect not only grammatical or semantic correctness, but also the elimination of any `non-native-like' stylistic patterns which hinder comprehension. For instance, native Spanish and Catalan speakers tend to construct overly complex sentences in English; hence, their manuscripts are often criticized and even rejected because of the excessively wordy phrasing or exceedingly awkward expressions. If EAL researchers were provided with specific strategies to minimize wordiness and avoid awkwardness, they might be able to enhance the readability of their manuscripts and increase the probability of success in the publication process. Given our interest in analyzing these complex areas of EAL production, we compiled a unique corpus of scientific manuscripts, written directly in English by UPV researchers and faculty, thoroughly revised by one of the present authors, and eventually published as peer-reviewed articles in English-language journals in their fields of study (e.g. thermodynamics, civil engineering, agricultural machinery, economics, biotechnology, crop production and food sciences). The initial corpus was created with 20 original manuscripts that included all the modifications written in by the linguistic consultant (author 3) together with the 20 published articles, which had been modified at the discretion of the researchers-authors. Each set of papers (manuscript draft(s) + published article) contained in the corpus was manually scrutinized by the linguistic analysts (authors 1 and 2), who assessed the differences between the original manuscripts and those accepted for publication. The initial analyses revealed a high frequency of reduction-type modifications, that is, many of the native consultant's suggestions targeted unnecessary, redundant and overly-complex phrases. Therefore, it seemed of interest to systematically identify the instances in the corpus and to classify what we call 'nip & tuck' procedures. These procedures aimed to effectively reduce (nip) the wordiness and rephrase (tuck) the awkwardness in the EAL production of these researchers-authors. In this paper, we shall first examine the unique features of this specific corpus and highlight the findings of the research conducted so far. Then, we will describe the corpus-based qualitative typology, developed from instances of wordy and awkward EAL writing patterns. Finally, we will conclude with suggestions as to how this typology may help Spanish researchers to improve their writing and broaden our understanding of the more complex processes involved in EAL production of scientific discourse.

Cruz-García, Laura and Heather Adams

Panel: 5. Corpus, estudios contrastivos y traducción

ADDRESSING THE POTENCIAL CUSTOMER IN FINANCIAL ADVERTS: A CONTRASTIVE ANALYSIS IN ENGLISH AND SPANISH

The aim of this study is twofold: (1) to identify and describe the linguistic resources that copy writers use in ads for financial products in order to establish the relationship between the addresser and the addressees in two different cultures (British and Spanish), and (2) to contrast the findings in each language and culture to find out to what extent this relationship differs from one language to another. To this end, we have analysed a corpus of 60 ads for financial products, made up of two sub-corpora (30 from the British and 30 from the Spanish mainstream press published in the first half of 2004) from both linguistic and pragmatic perspectives. The linguistic analyses carried out cover the most representative lexical, semantic, syntactical, graphic and phonic elements used to convey the advertising message, while the pragmatic analysis pays particular attention to the legal constraints pertaining in this product sector, as well as the role of consumer expectations, thus setting our linguistic analysis firmly within the social and cultural framework that gave rise to the production of these texts. Our analyses are carried out from the perspective of the translator's need to have a thorough knowledge of both the linguistic features and extra-linguistic factors that govern the production of a given type of text in a given cultural and communicative situation. Our intention is to explore and describe the differences that emerge from a detailed analysis of a representative sub-corpus in English and another in Spanish, each firmly embedded in their source culture. In order to determine the relationship existing between addresser

and addressee, we have looked at the register used in the texts, paying special attention to lexical and semantic elements such as the use of informal language, puns and figurative language, on the one hand; and morphosyntactic elements such as the personal pronouns and verb forms used by the addressers to refer to themselves and to the addressees. Our conclusions will be of interest not only to translators working in advertising but also to trainee translators (and their trainers), as pragmatic factors shape the forms of address used.

Cuenca, Maria Josep and Josep Ribera

Panel: 5. Corpus, estudios contrastivos y traducción

DEICTIC NEUTRALIZATION AND OVERMARKING IN TRANSLATING FICTION (ENGLISH-CATALAN)

Demonstratives, as space deictic elements, are analyzed in situational terms, that is, as linguistic items that point to elements of the situational ground of utterance with regard to the deictic origin. However, Corpus Analysis shows several puzzling facts from a traditional point of view: (i) non-situational uses outnumber the cases in which demonstratives indicate proximity or distance with respect to the addressor, (ii) non-situational demonstratives are frequently neutralized in translation (i.e., they are translated by a non deictic unit or deleted), and (iii) new demonstratives show up in the target text (that is what we call deictic overmarking). This research is based on a corpus of fiction in English and the translation of the texts into Catalan. The English demonstratives this/these and that/those and their Catalan counterparts have been analyzed and the general strategies activated in translation have been identified, namely: a) maintenance, b) shift, c) neutralization, and d) overmarking. In this presentation, neutralization and overmarking will be dealt with in detail. Our analysis puts forward that non-situational demonstratives are much more frequent in our corpus (400 cases, 83.5%) than situational ones and that neutralization is the most frequent strategy when translating them (177 cases, 44.3%). Non-situational deictics are frequently neutralized because they alternate with other phoric processes, such as ellipsis or 3rd person pronouns. In fact, Catalan shows a tendency to avoid deictic marking in syntactic contexts where the demonstrative could be interpreted as too focal or somehow emphatic. The strategy, which is mainly syntactically conditioned—neutralization is favoured when the demonstrative is in subject position or can be pronominalized by a clitic in the target language—, implies a loss of deictic force and sometimes also the empathetic nuance that the deictic adds, affecting the implication of the character or the narrator in the narration. On the other hand, overmarking is also very frequent, since many non-deictic English units are translated into Catalan by means of demonstratives (232 cases out of 519 demonstratives in Catalan, 44.7%). This translation strategy introduces in the target text subjective and intersubjective values not expressed in the source text. In conclusion, neutralization and overmarking are very frequent in translating fiction and have an effect on the target text by underspecifying or introducing, respectively, subjective and intersubjective values in the narration. The changes in the deictic perspective of the source text introduced by these strategies are not due to the systemic differences of the languages involved in the process of translation, but to syntactic and pragmatic factors leading to the underspecification or the introduction of the addressor's subjectivity in the target text.

Culy, Chris, Verena Lyding and Henrik Dittmann

Panel: 6. Corpus y variación lingüística

STRUCTURED PARALLEL COORDINATES: A VISUALIZATION FOR ANALYZING STRUCTURED LANGUAGE DATA

We present a visualization tool called Structured Parallel Coordinates (SPC), a specialization of Parallel Coordinates (cf., e.g., Inselberg, 2009), customized for the presentation and analysis of different types of structured language data, as found in corpora. We introduce three applications of the tool. They show SPC alone and as part of a broader process of data exploration, connected in particular with corpus queries. We provide detailed descriptions of the SPC visualizations and their interactive functionalities,

demonstrate how they can be employed in different linguistic analysis tasks, and explain the motivation behind design decisions taken to respond to characteristics of linguistic data. Parallel Coordinates are a way of representing multidimensional data using a two-dimensional display. Each dimension is represented along a vertical axis, and the values for a piece of data are connected by a line (see Figure 1). Interactive versions of Parallel Coordinates are flexible tools for data analysis, since selecting points and lines in the Parallel Coordinates display is the same as filtering the data (Inselberg, 2009). Parallel Coordinates are typically used with data dimensions that are conceptually independent, such as car size, year of manufacture, and mileage (cf. Frank and Asuncion 2010 for a standard test data set). However, language datasets often have dimensions which are interrelated or which have internal structure. One fundamental type of structure is the sequential order of linguistic units like words, phrases, or paragraphs. Another type of structure comes from meta-information associated with corpus texts, e.g. dates, where the data for each point in time can be treated as a dimension, and these dimensions are ordered (chronologically) with respect to each other. Rank orderings of (co-)occurrences of linguistic units provide an example of dimensions that have an internal structure: the ranks. SPC is designed specifically to deal with the special nature of structured language data such as these (cf. Collins et al. 2009 for another take on Parallel Coordinates for textual data). We present three applications of Structured Parallel Coordinates: (1) KWIC results as SPC, (2) ngrams and frequencies, and (3) ranking comparisons. Figure 1 shows a SPC display of the rank ordering by frequency of the top 20 (German) words starting with [Ss]elbst "self-", counted by lemma, in 5 years of newspaper text, ranging from 1991 to 2006. The words which do not appear in all years are grayed out, and the word Selbstbestimmung "self-determination" has been selected and highlighted with a thick line. The relative frequencies within years are indicated by green bars. SPC is a JavaScript tool that can easily be used with new kinds of data. For example, colleagues are using SPC to analyze learner texts. SPC and the applications are freely available under an Open Source license. SPC is an innovative tool for corpus analysis, which illustrates opportunities that are created when visualization techniques are adapted to the special needs of language information.

Currás Móstoles, Rosa and Miguel Angel Candel-Mora

Panel: 5. Corpus, estudios contrastivos y traducción

MÉTODOS DE LA LINGÜÍSTICA DE CORPUS APLICADOS A LOS ESTUDIOS DESCRIPTIVOS DE TRADUCCIÓN.

La comparación lingüístico-textual de un solo texto traducido con su original es una técnica reciente, que sin embargo, debe constituir la base imprescindible sobre la que realizar el comentario crítico de los textos traducidos y sacar conclusiones empíricamente fundamentadas acerca de lo que implica globalmente lo que llamamos traducción literaria. El análisis lingüístico-contrastivo entre pares de lenguas como base para la traducción tuvo su representación en las propuestas de autores como Catford, Vinay y Darbelnet, y en cuanto al par de lenguas inglés-español, Vázquez Ayora. El objetivo final en los Estudios de Traducción debería ser la conjunción de fuerzas entre las distintas disciplinas para contribuir a la práctica traductora y por consiguiente a su posterior análisis donde se valore su adecuación teniendo en cuenta la confluencia de circunstancias en el hecho traductor. El objetivo de este trabajo consiste en demostrar la consolidación del método de trabajo interdisciplinar, en este caso la por medio de la combinación de métodos de análisis procedentes de la Lingüística de Corpus y de los Estudios Descriptivos de Traducción para el análisis de traducciones teatrales. La primera parte del trabajo hace un breve recorrido por las particularidades del texto teatral y su traducción. En segundo lugar, se describe la metodología para la elaboración de un corpus específicamente orientado al estudio de una traducción teatral. Por último, partiendo de la observación empírica se presenta una clasificación de la problemática de la traducción teatral para posteriormente presentar las mejoras obtenidas tras la observación del corpus compilado ad hoc. Entre las características mas destacables cabe mencionar la adaptación de los métodos tradicionales de alineación al método más centrado en el teatro, así en lugar de alinear por segmentos de traducción se procede a alinear por réplicas, como unidad mínima de significado.

De Vos, Lien

Panel: 3. Estudios gramaticales basados en córpora

THE USE OF GENDER-MARKED PRONOUNS IN DUTCH: GRAMMATICAL VERSUS CONCEPTUAL GENDER.

The Dutch pronominal gender system provides a unique source for the investigation of variation and change, since it appears that the system is changing at a different pace within both varieties of Dutch. The northern Dutch variety, as spoken in the Netherlands, nowadays has a so-called semantic gender system: the choice of a particular pronoun depends on the conceptual properties of the referent, and no longer on the grammatical gender of the antecedent it refers to. The most crucial parameter in the process seems to be 'individuation': highly individuated nouns, such as count nouns referring to concrete entities, trigger the use of traditionally masculine pronouns, whereas lowly individuated nouns, such as abstract mass nouns, trigger the use of the traditionally neuter pronoun (Audring 2009). However, the Dutch gender system originally was grammatical, and gender-marked pronouns were strictly related to the grammatical gender of the antecedent noun. The southern Dutch variety, as spoken in Belgium, was believed to have retained this system in which pronouns agree in gender with their antecedent noun, which can be masculine, feminine or neuter. Recent studies have rendered this belief invalid, by illustrating that even adolescents do not yet reach an adult-like proficiency in the grammatical gender system, and that the influence of grammatical gender on pronominal reference gradually decreases from generation to generation (De Vos 2009, De Vogelaer & De Sutter to appear). Clear semantic patterns are observed, which may indicate the erosion of the original, grammatical system and the origination of a new, conceptually-based gender system. All of these previous investigations on southern Dutch have gathered their data in a similar way: by means of questionnaires, consisting of completion tasks. However, this excludes possible influence of discourse-factors on pronominal reference and it narrows down the view on semantic factors, since there is only a small amount of words under investigation. In this paper, these previous studies will be compared to a corpus-based investigation of the development in gender-marked pronouns in southern Dutch. The data is gathered from the Corpus Gesproken Nederlands 'Spoken Dutch Corpus', a nine million word corpus of contemporary spoken Dutch. The results of this paper will not only confirm the presence of semantic factors influencing the use of gender-marked pronouns, it will also supplement the existing data with a broader view on pronoun usage in spoken language. From these results it will follow that the choice between grammatical and conceptual (semantic) gender depends on much more than semantic factors, such as the discourse setting and linguistic context. The aim of it is to adjust and complement the ruling theories on this development of gender-marked pronouns in Dutch and to establish a framework that can be used for further research, which includes challenging some methodological issues.

Del Olmo Bañuelos, Elena, Antonio Moreno Ortiz and María Del Olmo Bañuelos

Panel: 8. Los córpora y la adquisición y enseñanza del lenguaje

COMPUTER LEARNER CORPUS (CLC) RESEARCH: UN FUTURO APOYO PARA MATERIALES DIDÁCTICOS BASADOS EN EL MÉTODO CLIL.

La aplicación de estudios basados en córpora en los campos de la Adquisición de Segundas Lenguas y de la Enseñanza de Lenguas Extranjeras lleva siendo patente desde los años ochenta. A nivel teórico, el uso de córpora en estudios lingüísticos ha permitido tanto la evaluación de teorías ya existentes, como la comprobación de nuevas hipótesis sobre lenguaje real; más aún, la sistematización y automatización que aportan estas herramientas de estudio puede proveer a las distintas ramas de la Lingüística Aplicada del rigor y consenso teórico necesarios para que se desarrollen como ciencias. A nivel práctico, el hecho de que el lenguaje real sea el objeto de estudio ha establecido una conexión más directa entre los que estudian la lengua y los que la enseñan (Granger, 2004, p. 123). La utilidad de los córpora de lenguaje real en el campo de la Adquisición de Segundas Lenguas así como en el campo de la Enseñanza de Lenguas Extranjeras no ha sido un descubrimiento reciente. Hoy día ya existen una multitud de materiales para la Adquisición y la Enseñanza del Inglés como Lengua Extranjera que se basan en muestras de lenguaje real: diccionarios de términos (Ehrlich, 1987), diccionarios de expresiones

idiomáticas (Deuter, Greenan, Noble, & Phillips, 2002), libros de texto (Hewings, 2005; McCarthy & O'Dell, 2004; 2005), etc. La diferencia que existe en este nuevo campo de estudio llamado Computer Learner Corpus (CLC) radica en la fuente de información: el aprendiz de esa lengua extranjera.

La ventaja que tienen los learner corpora es que permiten estudiar el uso de un idioma determinado por un hablante no nativo de forma cuantitativa. Muchos investigadores ya señalaron el potencial de los learner corpora en: a) el reconocimiento de etapas en el desarrollo de la interlengua (IL); b) estudios sobre transferencias lingüísticas derivadas de la primera lengua (L1); c) identificación del uso excesivo, o escaso, de determinados patrones lingüísticos; d) discernimiento de errores universales y errores radicados en L1; e) distinción entre el habla nativa y no nativa de una lengua. (Tono, 1999). En este trabajo exploramos la adecuación y aplicación de los CLC en entornos CLIL (Content and Language Integrated Learning), teniendo un objetivo doble: por una parte, repasaremos algunos de los ya existentes, exponiendo sus características principales. En segundo lugar trataremos de establecer una serie de criterios o pautas a seguir para su diseño y compilación, enfocadas a garantizar no sólo su reutilización, sino su explotación efectiva.

Díez Bedmar, María Belén

Panel: 8. Los córpora y la adquisición y enseñanza del lenguaje

SPANISH STUDENTS' MAIN PROBLEMS WHEN WRITING THE ENGLISH EXAM IN THE UNIVERSITY ENTRANCE EXAMINATION: A LEARNER CORPUS-BASED ANALYSIS

The research conducted on the English Exam in the University Entrance Examination in Spain has been divided into three main blocks (García Laborda, 2006): i) its validity design; ii) its construct validity, inter- and intra-rater reliability, the raters's scorings, etc; and iii) the need for the improvement of the exam. However, there have also been studies which have analysed the students' written production when taking this exam, as reflected in various (computer) learner corpora. In an edited book (Iglesias Rábade, 1999a), five papers presented the students' spelling errors (Doval Suárez, 1999), their morpho-syntactic errors (Crespo García, 1999), lexical errors (González Álvarez, 1999), problems in closed word classes (Woodward Smith, 1999), and in their textual organization (Iglesias Rábade, 1999b). Similarly, two PhD dissertations also focused on the students' errors when writing this exam in the foreign language by means of an Interlanguage Analysis (IA) or a Computer-aided Error Analysis (CEA). Thus, Wood Wood (2002) concentrated on the students' article use, and Rodríguez Aguado (2004) scrutinized their morphological and syntactic errors, as well as those problems related to orthography and vocabulary use. Despite the importance of these studies to know the main problems which pre-university students' show when writing in the foreign language, two main limitations can be found in these seven studies. First, each of them focused on a limited number of aspects of the foreign language, which results in an incomplete description of the students' written performance at this stage. Second, different methodologies were employed, e.g. various error-taxonomies, preventing the direct comparison of results. In order to bridge these two limitations, Díez-Bedmar (2010) analysed a representative sample of the compositions written on the same topic for the English Exam in the University Entrance Examination in Jaén in June 2008 by means of a CEA with the UCL Error Editor (Hutchinson, 1996), and the widely-used Error Tagging Manual, version 1.1. (Dagneaux, Denness, Granger and Meunier, 1996). This paper is divided into two main parts. The first one presents the findings obtained in Díez-Bedmar (2010), which allows an updated description of the students' profile at this stage of their foreign language acquisition process. The use of a widely-used error taxonomy also entails the comparison of results with those provided in the extensive research which has also employed the Error Tagging Manual in the Spanish and international contexts. In the second part of the paper, a comparison is made between the findings in Díez-Bedmar (2010) and those presented in the above-mentioned publications, so that it is possible to point to interesting tendencies regarding the common errors made by secondary-school leavers. The information offered in this paper may prove the starting point to cater for the students' empirically-based needs at this stage, by means of teaching materials at the end of the secondary school education, or the design of appropriate courses when entering the European Higher Education Area (EHEA) in Spain.

Duran, Isabel

Panel: 1. Diseño, compilación y tipos de córpora

CRITERIOS ESPECÍFICOS PARA LA ELABORACIÓN Y DISEÑO DE LOS CORPUS ESPECIALIZADOS PARA LA TERMINOGRAFÍA

La especificidad de la Terminografía basada en corpus (Meyer y Mackintosh, 1996: 258), en contraposición a la Lexicografía basada en corpus u otras aplicaciones de los corpus (traducción, enseñanza de segundas lenguas, etc.), obliga al establecimiento de una serie de requisitos o criterios específicos para el trabajo terminográfico. Algunos de ellos serán comunes a los criterios generales de la compilación y diseño de los corpus y otros, como veremos, presentarán algunas diferencias. Antes de comenzar con los criterios específicos, consideramos necesario exponer las fases en las que se divide el trabajo de un terminógrafo, con objeto de indicar explícitamente las necesidades del empleo de los corpus en cada fase y, así, poner de relieve la importancia de la compilación de los corpus: en primer lugar, los terminógrafos deben familiarizarse con el dominio en el que están trabajando, a fin de establecer sus límites, relaciones con otros dominios y la organización interna de este, es decir, los subdominios en los que puede dividirse; en segundo lugar, deben identificar las fuentes de conocimiento que les proporcionarán tanto la información lingüística como conceptual y comunicativa de los términos; en tercer lugar, los terminógrafos pasan a considerar, con la ayuda de las fuentes de conocimiento, un conjunto de candidatos a términos para empezar a trabajar y crear la conceptualización del dominio; en cuarto lugar, analizan la nomenclatura identificada en la fase previa, así como la información terminológica de los textos compilados (colocaciones, relaciones semánticas, etc.) y elaboran la base de datos (onto)terminográfica extrayendo del corpus información para realizar definiciones y seleccionar contextos adecuados; por último, resuelven los posibles problemas presentados, realizan las validaciones y editan el recurso terminológico. En estas fases, se observa la importancia que tiene la documentación en la labor terminográfica y, por ende, la relevancia que presentan los corpus electrónicos para el terminógrafo durante todo el trabajo. Partiendo de esta situación, podremos determinar cuáles son los criterios que se deberían seguir a la hora de compilar un corpus especializado para tareas terminográficas. Por un lado, nos encontraremos criterios generales concretados según las necesidades de los usuarios, en este caso los terminógrafos, como son el criterio de la cantidad, el criterio de calidad, el criterio de simplicidad (referido a la cantidad y al tipo de información añadido al texto original) y el criterio de documentación. Además de estos criterios generales, consideramos que son útiles otros, de carácter más específico, aunque muy relacionados con los anteriores, a saber: delimitación clara del campo de trabajo y, por ende, del corpus; criterio de apertura del corpus y el criterio del medio de producción del texto (oral o escrito). En nuestra opinión, estos serían los criterios básicos que debería cumplir cualquier corpus especializado que se utilizara para cualquier tarea terminográfica. A lo largo del trabajo, se desarrollarán estos criterios y se realizarán comparaciones con la aplicación de estos criterios en otras disciplinas, como puede ser la lexicografía, la enseñanza de la traducción, etc.

Ekaterina Tarpomanova, Svetlozara Leseva, Svetla Koeva, Borislav Rizov, Hristina Kukova, Tsvetana Dimitrova and Maria Todorova

Panel: 1. Diseño, compilación y tipos de córpora

DESIGN AND DEVELOPMENT OF THE BULGARIAN SENSE-ANNOTATED CORPUS

The paper describes the methodology, compilation, annotation and applications of the Bulgarian Sense-Annotated Corpus (BulSemCor) - a manually annotated corpus of over 100,000 words in which each language unit (LU) is assigned a sense according to the Bulgarian wordnet (BulNet). The input corpus is an excerpt from a general structured corpus of contemporary Bulgarian designed according to the Brown Corpus methodology. The input corpus consists of over 800 text units of 100+ words each, selected according to the density of highest frequency open-class lemmas. The corpus is represented in

a flat xml format. The text is encoded as a list of xml tags 'word' whose attributes store relevant information such as form, lemma, selected sense, annotator. Another attribute encodes a parent ID that links the tokens identified as part of a compound. The corpus annotation tool provides a number of functionalities such as (i) input data editing including insertion and deletion of tokens, identification of MWEs with contiguous or non-contiguous constituents; (ii) flexible text navigation strategies - forward and backward navigation according to a given criterion such as all words, non-annotated words, all instances of a current sense or word, etc.; (iii) flexible search strategy allowing both exact match search according to wordform or lemma, and regular expression search. The tool interface features fully-fledged visualisation of the wordnet synsets for the available candidate senses for a selected LU through coupling with the system for wordnet development and exploration. The annotation tool is OS independent, adaptable to annotation schemes for different language levels, affords multiple-user concurrent access and dynamic real time update of changes in the knowledge base. The annotation of BulSemCor involves the following steps. In the preprocessing stage automatic lemmatization is performed. Next, the LUs are mapped to the corresponding BulNet senses through their lemma. The semantic annotation proper consists in the selection of the correct sense from the available candidates. The annotated LU inherits all the information contained in the selected synset, thus receiving morpho-syntactic annotation (through the POS) besides the semantic one. One of BulSemCor's main features is the exhaustive annotation approach requiring that each LU be annotated. It has resulted in: (i) enlargement of the Bulgarian wordnet with closed-class words and language specific concepts; (ii) reconsideration of a number of theoretical assumptions; (iii) practical decisions regarding interlingual asymmetry. The main application of BulSemCor is to serve as a training corpus for WSD tasks. It has already been employed in two implementations. In the first one based on Hidden Markov Models, BulSemCor has been used in the training and evaluation. A second, knowledge-based implementation currently under development, uses it mainly for the purposes of evaluation. BulSemCor has a variety of applications in linguistic research from lexicology and lexicography to semantics, grammar, stylistics, etc. An online demo of the corpus has been implemented and made publicly available. It affords search for words according to wordform or lemma. The available senses are sorted according to frequency of occurrence and are supplied with a gloss and an example.

Enghels, Renata and Marlies Jansegers

Panel: 5. Corpus, estudios contrastivos y traducción

HACIA UN ENFOQUE EMPÍRICO EN LA SEMÁNTICA: EL PAPEL DE LA TRADUCCIÓN. ESTUDIO CONTRASTIVO DEL VERBO SENTIR.

Es bien sabido que en las últimas décadas, la lingüística de corpus se ha revelado útil tanto para el estudio de aspectos morfosintácticos de la lengua, como para estudios de índole semántica (cf. entre otros Geeraerts 2010, Glynn 2010, Oster, 2010). Sin embargo, el uso de corpus para análisis semánticos conlleva ciertas dificultades metodológicas. Así, se distinguen por ejemplo distintas formas para recoger los datos, para analizarlos y además, una amplia gama de técnicas cuantitativas para el tratamiento de los resultados (cf. Glynn 2010). La presente investigación se inscribe esencialmente en la primera problemática, o sea la colección y la elección de los datos que se presten a un análisis semántico detenido. De hecho, el objetivo principal de esta ponencia consiste en examinar en qué medida dos tipos distintos de corpus (un corpus paralelo y otro comparable) pueden inducir a resultados complementarios para la investigación semántica. Nos preguntamos más particularmente hasta qué punto ambos tipos pueden ser complementarios para la caracterización y el análisis de los llamados 'cuasi sinónimos' entre lenguas. Con este objetivo, se presentará el estudio de caso concreto de un verbo español y sus cognados en otras lenguas romances como el francés y el italiano. Más precisamente, este estudio examina si los verbos sentir en español, francés e italiano – además de ser cognados morfológicos – pueden considerarse también verbos cognados desde una perspectiva semántica. De hecho, los significados diferentes de sentīre en latín surgen en las diferentes lenguas románicas, pero se nota que el verbo ha sufrido al mismo tiempo una especialización semántica. En francés, ésta se sitúa más bien en el campo de la percepción olfativa (pej. sentir l'odeur des cuisines) (cf. Franckel/Lebaud 1995) y – relacionado con esta percepción física – el campo de la cognición (je sentais

que le monde était plus complexe que nos discours). El italiano, en cambio, ha optado claramente por la percepción auditiva, tanto en su uso activo (pej. sentir la ràdio) como pasivo (pej. sentir la voce) (Badynska-Lipowczan 1996) y utiliza el verbo incluso en actos de comunicación (Emma Marcegaglia ha sentito (e sentirà) i vertici Fiat) o como interjección (senti, tu lo conosci un tale che si chiama Angelo Pardo[…]). Finalmente, en español predomina la percepción emotiva del verbo y – vinculado a esta percepción subjetiva – el significado particular de 'lamentar' ('lo siento').

Ezeiza, Joseba

Panel: 8. Los córpora y la adquisición y enseñanza del lenguaje

PLATAFORMA GARALEX: INFRAESTRUCTURA TECNOLÓGICA PARA LA INVESTIGACIÓN Y LA DIDÁCTICA DE LENGUAJE DEL ÁMBITO DE LAS CIENCIAS JURÍDICAS

En esta comunicación vamos a presentar un proyecto de una estructura tecnológica administrada en Web, para la investigación y la didáctica del lenguaje jurídico en sus diversos registros y niveles aplicando metodologías de análisis de corpus (Monzó y Borja, 2000 y 2001; Bowker, L. & Pearson, J., 2002; Elosegi, 2002; Biber, 2006; Lersundi et al., 2008; Parodi, 2007; Ezeiza, 2008 y 2009; Taylor et al., 2008; Zabala et al., 2008; Lombardo, 2009; Azkarate, 2009). Se trata de una Web en construcción (fecha prevista de lanzamiento: marzo 2011) que ofrecerá a estudiantes, profesores y profesionales del ámbito jurídico y administrativo tres tipos de recursos: a) recursos de comunicación; b) recursos de consulta; y c) recursos de formación. Se trata de un proyecto que tiene como finalidad contribuir a dinamizar y armonizar el desarrollo y el uso de la lengua vasca entre los especialistas del área jurídica en el entorno universitario: profesores, estudiantes, investigadores, etc. El núcleo central de la plataforma lo ocupará un "taller" para el análisis de las producciones académicas y profesionales. Dicho taller contará con varios módulos: a) una base documental de textos especializados; b) un instrumento de análisis de corpus; c) un extractor terminológico; d) una base de datos terminológica; y e) una base de datos de fichas de estilo. El taller está pensado para facilitar la colaboración entre lingüistas y expertos en Ciencias Jurídicas y tiene como objetivo principal ofrecer información relevante para la investigación y la enseñanza del lenguaje jurídico en la universidad, tanto a estudiantes como a profesores.

Ezeiza, Joseba and Agurtzane Elordui

Panel: 1. Diseño, compilación y tipos de córpora

HERRAMIENTAS Y CRITERIOS PARA LA CREACIÓN DE UN BANCO DE CONOCIMIENTO SOBRE LOS USOS DEL LENGUAJE EN LA RED

Las nuevas formas y modos de comunicación en la red han generado un renovado interés de los lingüistas y otros profesionales por los usos del lenguaje en los nuevos canales y soportes (Díaz Noci, 2001; Ferris, 2002; Hasan & Martin, 2002; Machón, 2003; García, 2005; Lamarca, 2006; Canhavilas, 2007; Franco, 2009; Yus, 2010). Una de las líneas de trabajo que promete ser productiva en este ámbito es la basada en el estudio de corpus (Reppen et al., 2002; Lim et al., 2004; Biber & Kurjian, 2006; Hund eta al., 2007; Meyer & Stein, 2009; Renouf. & Kehoe 2009; Aguado de Cea et al., 2010). En esta línea, un equipo formado por investigadores de la UPV/EHU y del centro de investigación Ametzagaiña I+D, ha desarrollado una plataforma concebida para la creación de un banco de conocimiento sobre los usos lingüísticos en la red basado en estudios de corpus. Dicha plataforma consta de cuatro bases de datos integradas que pueden gestionarse desde la web: una base de datos bibliográfica, una base de datos documental, una de usos léxicos y una sobre cuestiones de estilo. La base de datos documental permite generar y gestionar de manera flexible un número indeterminado de corpus de documentos textuales, hipertextuales, multimedia e hipermedia en diversos formatos (pdf, doc, HTML, jpg, mp3, etc.) y, gracias a la estructura taxonómica facetada sobre la que opera, facilita una caracterización muy precisa de los documentos alojados en ella. Para ello cuenta con varios instrumentos que dan la posibilidad de obtener información lingüística relevante, entre los que destacan un motor de búsqueda por lemas o palabras,

uno de búsqueda de cadenas de hasta cinco lemas o palabras, otro de categorías morfológicas, un instrumento de cálculo y comparativa de frecuencias de uso, un instrumento de búsqueda de combinaciones léxicas de dos y tres elementos, un motor de búsqueda de patrones sintácticos y una herramienta para discriminar el léxico más representativo. Todas estas herramientas operan bien sobre el corpus en su conjunto o bien sobre una determinada selección de documentos que compartan rasgos contextuales (ámbito de producción, modalidad de comunicación, interlocución…), temáticos (tema, subtema, tratamiento del tema…), funcionales (tipo de documento, género, subgénero…) o estructurales (superestructura, macroestructura, microestructura…) o cualquier combinación de rasgos que se considere pertinente. Ello hace posible la realización de análisis estratificados y comparativos muy detallados. En la versión piloto, esta infraestructura aloja un corpus de documentos del ámbito del (ciber)periodismo y otro corpus de documentos del ámbito de la (ciber)literatura. Cada uno de ellos cuenta con un desarrollo específico de la estructura taxonómica básica sobre la que opera la base documental. Actualmente la plataforma opera únicamente en lengua vasca, pero no se descarta abordar en el futuro el desarrollo de una versión multilingüe. En cualquier caso, la estructura taxonómica es independiente de esta variable y puede ser transferida (de forma integral o parcial) a cualquier otra herramienta, lengua o proyecto que esté interesado por el estudio de los rasgos lingüísticos de la cibercomunicación.

Faya Cerqueiro, Fatima

Panel: 6. Corpus y variación lingüística

REQUEST MARKERS IN DRAMA: DATA FROM THE CORPUS OF IRISH ENGLISH

In the Late Modern English period we observe a change in the use of main request markers, whereas pray was the most common courtesy marker in requests at the beginning of this period, it was eventually replaced by please and the former marker disappeared entirely in the twentieth century. A preliminary study in ARCHER (A Representative Corpus of Historical English Registers) showed that these markers were found mainly in three types of texts, namely letters, fiction and drama. The analysis of those items in novels and letters have already brought interesting results about the evolution of these markers and especially about the replacement of pray by please (cf. Faya Cerqueiro 2008 and 2009). Nevertheless, requests markers have not been studied in drama texts yet. Therefore, an analysis of plays will help to complete the whole picture of the main request markers in the Late Modern English period and will allow text-type comparisons. For this purpose I will make use of the Corpus of Irish English. The Corpus of Irish English collects Irish documents written in English from the early fourteenth century up to the twentieth century, allowing diachronic analyses. The different genres represented in this corpus comprise poetry, glossaries, sketches and full-length plays, although drama is the best represented genre in the corpus. The material compiled from the sixteenth to the eighteenth centuries in the corpus includes not only “genuine representations of Irish English by native Irish writers” but also “texts by non-Irish writers where the non-native perception of the Irish English is found” (Hickey 2003: 242). As regards number of words, the drama selection of this corpus contains an approximate number of 500,000 words, although the twentieth century provides almost half of them. Drama is probably the most profitable fictional genre for the study of pragmatic issues, especially those regarded as typical of the spoken language. Even though it should be admitted that this genre contains an imitation of actual speech, it represents the spoken medium as close as possible and if it is “used with the necessary caution, plays may also yield insights into what counted as polite or impolite behaviour and how, for instance, greetings, insults or compliments were realised at that time” (Jucker 1994: 535). Culpeper and Kytö (1999) classify drama as constructed dialogue with minimum of narratorial intervention, since apart from stage directions, plays contain dialogue almost exclusively. There are important contributions to historical pragmatics using only drama, proving the relevance of this text-type in pragmatic analysis (cf. Brown and Gilman 1989).

Fernández-Villanueva Jané, Marta and Oliver Strunk

Panel: 5. Corpus, estudios contrastivos y traducción

CONECTORES CAUSALES EN LA LENGUA ORAL. UN ANÁLISIS CONTRASTIVO BASADO EN CORPUS ENTRE ALEMÁN Y CATALÁN.

Los conectores causales se manifiestan de forma explícita dentro de los textos, y esto los convierte en un buen indicador de las estrategias utilizadas por el hablante para establecer relaciones causales entre proposiciones, pues a diferencia de otras estrategias comunicativas, es uno de los recursos lingüísticos claramente identificables con los métodos de la lingüística de corpus. Por medio del uso de dos corpus estructuralmente comparables, uno en alemán nativo, el otro en alemán realizado por aprendientes, analizaremos las potenciales diferencias en el uso de conectores textuales a nivel inter- e intratextual, y si estas potenciales diferencias pueden relacionarse con otras variables independientes. Las divergencias en el uso pueden integrarse finalmente en un sistema de indicadores complejo para determinar el nivel de lengua de un aprendiente de alemán como lengua extranjera, que no es objeto de discusión. Los corpus usados son los que se han elaborado en el marco del proyecto Varkom (Fernández-Villanueva, Strunk 2009). Incluyen las transcripciones de entrevistas estructuradas segmentadas según tipos textuales y tienen una base de informantes comparable.

Fragaki, Georgia

Panel: 2. Discurso, análisis literario y corpus

EVALUATIVE ADJECTIVES IN A CORPUS OF GREEK OPINION ARTICLES

Existing attempts to describe evaluation in text treat adjectives as mere devices of evaluation. However, the reverse question has not been raised: which are the adjectives that can function evaluatively in texts? The answer commonly given to this is descriptive adjectives (cf. Hewings 2004: 253) or adjectives having positive or negative meaning, relative or superlative degree, or gradability, that is having the typical features of descriptive adjectives (cf. Hunston & Francis 2000: 188-189, Hunston & Sinclair 2000: 91). A systematic corpus-based study of adjectives can reveal a different picture: Fragaki (2010) claims for Greek that several adjective categories can assume an evaluative function, among which a special category of evaluative adjectives, whose exclusive function is evaluation. The aim of this paper is to contribute to the description of the category of evaluative adjectives, drawing on a corpus of opinion articles from the Corpus of Greek Texts (CGT), a reference corpus of Greek. The corpus of the study includes texts of 450,576 words from three Greek newspapers of different political orientation. It is suggested that, while descriptive adjectives are commonly used for the attribution of a good or a bad property to an object of evaluation, the category of evaluative adjectives is used for evaluation relating to modality, comment, intensification and importance. With respect to these functions, four groups of evaluative adjectives are distinguished: a) modal adjectives, b) comment adjectives, c) intensifying adjectives and d) adjectives of importance. The criteria used for this classification are both functional and semantic and are based on extensive corpus analysis of the data. It is notable that two of these groups (modal adjectives and adjectives of importance) concur with Hunston's (1994) and Thompson & Hunston's (2000) parameters of evaluation. In addition, modal adjectives as carriers of deontic or epistemic modality, as well as intensifying and adjectives of importance as a means of denoting the degree to which something happens or the importance with which something is viewed, contribute indirectly to the positive or negative evaluative frame of the text (cf. attitudinal frame, Bublitz 2002). Finally, comment adjectives are employed for making a (usually) negative comment on an object of evaluation and in this way offer direct evidence for the evaluative frame of the text.

Frías Delgado, Antonio

Panel: 7. Lingüística computacional basada en corpus

ESTUDIO COMPARATIVO DE COLOCACIONES EN TEXTOS ORIGINALES Y EN SU TRADUCCIÓN

En el trabajo se comparan las colocaciones en textos originales (ruso, inglés, francés, alemán, español, italiano) con las de sus traducciones (español, inglés, alemán, italiano, francés). Se usan las técnicas habituales en Lingüística Computacional. Los resultados muestran una fuerte discrepancia entre ambas listas que no son imputables meramente a las características de ambas lenguas.

Froehlich, Heather

Panel: 1. Diseño, compilación y tipos de córpora

ARE YOU A MAN?: ON SEEING GENDER IN SHAKESPEARE

Through a literary-linguistic, discourse-oriented computational approach I will present a new way to find patterns of gender in Early Modern drama. Building on previous corpus stylistic studies (Culpeper 2001 and 2002, Hunston and Francis 2000, and Fischer-Starke 2010), I suggest that the use of gender-specific terms are not in proportion to the character population of a play. Using AlphaX and Excel, I assemble examples of both grammatical gender and natural gender within the context of a line of Shakespeare's plays. This study presents a comprehensive overview of grammatical (subject/object) and thematic roles through a comparative study of third-person personal pronouns and gender-specific nouns in Macbeth and The Merry Wives of Windsor through the building of a pilot database of each word within the context of a sentence. The relationship of grammatical and semantic roles are encoded and thus manifest themselves into a literary representation of gender: the textual representation of gender is encoded by the language used. Macbeth is a play that is very concerned with masculinity, whereas The Merry Wives of Windsor focuses primarily on women. Gender identification in both plays in proportion to the gender representation of characters is less overt and more often encoded in the text itself: through the building of this database, I comment on the predictability of gender representation in relationship to the gender proportions of a cast. The implications of proportional representation of a cast have been largely ignored in (feminist) stylistic studies of Shakespeare's texts, a field which chooses instead to focus on the overt patriarchical structures presented in Early Modern drama; my study begins to fill this void through a critique of Shakespeare's plays as a (proto)feminist texts.

Fuster Márquez, Miguel and Begoña Clavel Arroitia

Panel: 8. Los córpora y la adquisición y enseñanza del lenguaje

ENGLISH LANGUAGE TEACHING AND LEARNING IN TERTIARY EDUCATION: CORPUS CHOICE AND IMPLEMENTATION

The aim of this contribution is to propose a model to integrate corpus linguistics (CL) in the teaching of the English Language at university level. This research is still in progress since we need to assess the results at the end of this academic year. The subjects of this study are our own students in the second year of the compulsory module (English Language IV) of the newly implemented degree of English Studies at the Universitat de València. It is precisely in the new university paradigm, in which students are required to learn to learn, develop skills and solve problems autonomously, that the deployment of corpus methodologies contributes to the enhancement of students' potential in such a direction. As Sinclair (2004) points out, students should be given the opportunity of consulting authentic language and corpus-based methodologies may come to cater for that need. It remains true that after decades of CL, even those textbooks targetting advanced learners contain written and spoken language samples which are not authentic. Exclusive exposure to textbooks cannot be sufficient if we wish our students to grasp more fully how real language actually works. Our study focuses specifically on the development of writing which fits in with the long tradition of corpus research devoted to productive written skills. It is our contention that if teachers are willing to embark on this type of experience there is no need to resort to large reference corpora, such as the BNC or the COCA, or The Bank of English although these are truly invaluable sources. However, a much modest proposal would consist in compiling smaller corpora which can immediately be applied offline in the classroom through freeware tools such as AntConc. Our proposal is structured around three corpora. The first corpus we have designed contains

updated articles of leading newspapers from the UK and the USA, which have been gathered by means of Lexis Nexis. This corpus can be used when what we have in mind is "general English". A second corpus contains recent academic articles published in leading journals, but exclusively in the field of humanities. This corpus meets the demands of the curriculum in our degree, since our students' learning goals include the attainment of competence in academic English in the field of the Humanities. And the third one is a much more modest ad hoc learner corpus which contains our own students production, with the hope of obtaining a much more accurate picture of their learning stage. The aim of this whole project is no other than to offer a coherent procedure to promote corpus exploitation, either indirectly by teachers through the design of corpus-based activities, or through hands-on corpus exploration by students. We believe that an inductive approach through corpus-driven awareness-raising activities is in conformity with the main guidelines being implemented in higher education pedagogy.

Gallego-Hernández, Daniel and Ramesh Krishnamurthy

Panel: 5. Corpus, estudios contrastivos y traducción

COMENEGO (CORPUS MULTILINGÜE DE ECONOMÍA Y NEGOCIOS) VS. METODOLOGÍAS WEB AS/FOR CORPUS APLICADAS A LA PRÁCTICA DE LA TRADUCCIÓN ECONÓMICA, COMERCIAL Y FINANCIERA

La práctica de la traducción económica, comercial y financiera requiere especialmente el desarrollo de la competencia instrumental o documental de los traductores en formación que permita suplir posibles carencias de conocimientos especializados. Esta competencia tecnológica implica el uso no solo de fuentes lingüísticas, como bases de datos terminológicas, sino también de fuentes textuales, como textos paralelos, es decir, textos comparables respecto de la función, el tema o la situación comunicativa de los textos originales objeto de traducción. Las actuales posibilidades tecnológicas han llevado a asociar el empleo de textos paralelos en traducción especializada a la explotación de corpus. En este sentido, los corpus se conciben como un de conjunto de textos paralelos del que el traductor puede sacar provecho (extracción de terminología, búsqueda de paralelismos conceptuales, análisis discursivo, etc.). En el mejor de los casos, estos recursos lingüísticos pueden ya estar compilados y disponibles en Internet. En cambio, si el traductor de textos especializados se enfrenta a un texto cuyo campo de especialidad no se encuentra entre los recursos textuales de los corpus disponibles en Internet, es él mismo quien puede compilar su propio corpus ad hoc (Corpas Pastor, 2001, 2004; Sánchez Gijón, 2002, 2004, entre otros). En el caso de la traducción económica, comercial y financiera francés-español y español-francés, existen en la actualidad pocos corpus virtuales que puedan servir de apoyo a la práctica de este tipo de traducción: el corpus técnico del IULA, aunque es de libre acceso, solo contiene un subcorpus español de economía de alrededor de un millón de palabras; CLUVI permite consultar textos sobre economía y consumo en español, además de otras lenguas, pero ninguna francesa; el MLCC Multilingual and Parallel Corpora contiene un subcorpus genérico de artículos financieros de periódicos en francés y español, pero es de pago; Vicente (en prensa) posee un corpus representativo del lenguaje especializado del comercio electrónico en la prensa general y especializada en francés y español, pero es privado. Ante este panorama, como formadores de traductores para el ámbito de la economía y los negocios, nos vemos obligados actualmente a implementar en el aula dos tipos de metodologías de explotación de textos paralelos: una que considera la web como si fuera un corpus (web as corpus), con la que el traductor utiliza los buscadores como si fueran herramientas de concordancias (Gallego Hernández, 2010a); y otra que emplea la web para compilar corpus (web for corpus) y que además requiere el desarrollo de una competencia instrumental relacionada con conocimientos informáticos (Gallego Hernández, 2010b). COMENEGO está pensado, entre otras cosas, para que el traductor en formación no invierta tanto tiempo en la búsqueda de textos o en la compilación ad hoc de corpus y pueda dedicarse directamente a sacar provecho de las funcionalidades típicas de los textos paralelos. En este artículo trataremos los temas relacionados con el diseño y la creación de este corpus, así como sus ventajas y desventajas que, en un futuro, pensamos que puede presentar respecto de metodologías ad hoc para la explotación de textos paralelos.

Gallego-Hernández, Daniel and Miguel Tolosa-Igualada

Panel: 5. Corpus, estudios contrastivos y traducción

ELABORACIÓN DE GLOSARIOS A PARTIR DE CORPUS PARALELOS AD HOC. APLICACIÓN A LA INTERPRETACIÓN DE CONFERENCIAS EN EL ÁMBITO SOCIOECONÓMICO

La interpretación de conferencias es una actividad que, dadas las condiciones espacio-temporales en las que se desarrolla, no da pie, a diferencia de la traducción escrita, a que los profesionales que se dedican a ella puedan documentarse, al menos no de manera exhaustiva, durante el proceso de escucha activa-reformulación. Considerando, por otra parte, que el intérprete debe estar, en principio, dispuesto a aceptar cualquier encargo, independientemente del tema principal de las conferencias y de los conocimientos previos que tenga sobre este, el trabajo documental inherente a la preparación de cualquier interpretación deberá llevarse a cabo antes de su celebración. En la actualidad, gracias al desarrollo que han experimentado las tecnologías de la información y la comunicación durante estos últimos años, las labores de documentación tienden a asociarse, entre otras cosas, al trabajo con corpus, especialmente en traducción de textos especializados. Ello implica que el traductor puede compilar en su ordenador los textos paralelos a los que consigue acceder en la web y explotarlos mediante las aplicaciones informáticas de gestión de corpus disponibles, para satisfacer las necesidades informativas que van surgiéndole durante la actividad traslativa. El intérprete, por su parte, tiene la posibilidad de obrar igualmente con el objetivo básico de extraer de este tipo de recursos textuales, en forma de glosarios, el vocabulario de sus lenguas de trabajo referido al tema principal de la conferencia en la que se han solicitado sus servicios, etc. y anticiparse así, en la medida de lo posible y razonable, a los eventuales problemas y dificultades que puedan presentársele durante la interpretación. En este trabajo nos proponemos reflexionar sobre los pasos que, en la etapa de documentación previa a la conferencia, el intérprete puede dar para elaborar este tipo de glosarios a partir de corpus paralelos compilados ad hoc, haciendo uso de las aplicaciones gratuitas disponibles en internet.

Garazi Olaziregi, Francisco Javier Calle and Dolores Cuadra Fernández

Panel: 7. Lingüística computacional basada en corpus

COGNOS TOOLKIT: UN CONJUNTO DE HERRAMIENTAS PARA LA ANOTACIÓN LINGÜÍSTICA DE CORPUS

En esta comunicación se presenta un conjunto de herramientas para el análisis integral de corpus, reunidas bajo la denominación común Cognos Toolkit. Asimismo, también se exponen los resultados obtenidos en su utilización en la construcción de sistemas de interacción multimodales enmarcada en distintos proyectos de investigación de financiación nacional. Cognos incluye desde herramientas metodológicas para el análisis integral de corpus afectando a distintos ámbitos (aunque prestando mayor atención a los aspectos lingüísticos), hasta aplicaciones software que facilitan ese análisis, llegando incluso a automatizar alguno de los procesos más mecánicos. En ese abanico se incluye también un lenguaje de formalización de las anotaciones realizadas mediante la herramienta, soportado por un esquema XML, que posibilita la reutilización y compartición de los corpus anotados. Aunque sujetas a evolución, algunas de estas herramientas ya han sido publicadas mediante licencia GNU para su uso gratuito. Son las de corte más lingüístico, y se restringen a la interpretación y generación de lenguaje natural y el análisis pragmático de diálogos (Cognos.CA, Cognos.NL y Cognos.DIAL). A medida que se emplean las aplicaciones anteriores, además de las anotaciones recogidas en un fichero de acuerdo con el esquema definido, se alimenta una base de conocimiento común para las tres aplicaciones, cohesionándolas de esta forma. El conocimiento almacenado es accesible para las tres aplicaciones, de manera que facilita –e incluso automatiza en algunos aspectos- el proceso de anotación de las muestras posteriores. Aunque la metodología define un orden para las distintas fases en las que se emplea cada aplicación, éstas pueden emplearse en cualquier momento para consultar e incluso actualizar la base de conocimiento. La primera de las aplicaciones, Cognos.CA, permite definir un conjunto de actos comunicativos que podrán vincularse a tantos corpus como se desee. Estos conjuntos de actos son creados de acuerdo con una taxonomía definida en la metodología, de manera que el procedimiento de creación de nuevos actos comunicativos requiera que los usuarios definan los

parámetros necesarios para ubicar el nuevo acto en la misma. Aunque el usuario no conozca los fundamentos teóricos de la taxonomía, ni tan siquiera la existencia de la misma, la herramienta clasifica automáticamente los actos, minimizando las ambigüedades y solapamientos entre los elementos de cada clase, obteniendo conjuntos estandarizados, combinables y reutilizables. Además, proporciona los mecanismos necesarios para concordar los elementos de dos conjuntos de actos cualesquiera, posibilitando la traducción de un corpus anotado sobre un conjunto a otro, y extendiendo así sus posibilidades de reutilización. Cognos.NL es la aplicación que permite a los analistas generar patrones (gramáticas relajadas) partiendo de expresiones en lenguaje natural, asociando a cada uno de ellos uno o varios actos comunicativos definidos previamente con la aplicación Cognos.CA. El resultado de esta fase es complementario al que se obtiene mediante Cognos.DIAL. Por último, la aplicación Cognos.DIAL, formada por distintos módulos complementarios. El primero de ellos, Cognos.DIAL.Indiv, permite anotar cada diálogo del corpus de forma independiente al resto. El siguiente módulo, Congos.DIAL.Global, facilita la identificación de diálogos equivalentes (y todas sus alternativas) unificando dos o más muestras anotadas con Cognos.DIAL.Indiv.

García Varela, Ana Patricia

Panel: 5. Corpus, estudios contrastivos y traducción

'WHEN POLICE ARRIVED AT THE SCENE' OR 'HAN VENIDO DOS POLICÍAS': ON THEME AND THEMATIC PROGRESSION IN NEWS REPORTS*

In this paper I shall explore the interaction between Theme-Rheme choices across English and Spanish journalistic discourse in order to see how this interaction is instantiated in the two languages (Halliday & Hasan 1976; Halliday 1985; Francis 1989, 1990; Fries 1994; Gómez-González 1994, 2001; Taboada 1995; Halliday & Mathiessen 2004; Arús Hita 2010). In particular, two research questions will be addressed:

1) Which Theme-Rheme patterns characterize journalistic discourse in English and Spanish?

2) Which patterns of Thematic Progression are more recurrent in this genre across the two languages?

The data will consist on news reports dealing with cases of domestic violence extracted from the online versions of four journals: The Guardian and The Times (English), on the one hand, and El País and El Mundo (Spanish), on the other. The results show that, despite the typological differences between English and Spanish, the thematic organization of news reports is, in general terms, rather similar in the two languages, although differences in the length of news reports as well as in the thematised elements are salient.

Garcia-Pastor, Maria Dolores

Panel: 8. Los córpora y la adquisición y enseñanza del lenguaje

LEARNERS' DISAGREEMENTS IN EFL: L2 PRAGMATICS AND THE USE OF A LEARNER CORPUS IN THE LANGUAGE CLASSROOM

The instruction and learning of pragmatic issues in a second or foreign language (L2 pragmatics henceforth) has been granted increasing attention recently as reflected in current European trends that search for innovation and development in second/foreign language (L2/FL) teaching and learning (García-Pastor, 2009, in press). Likewise, the use of corpora in English language teaching (ELT) has been encouraged in the past few years in an attempt to foster new advances in the field (cf., Bellés-Fortuño et al., 2010). This study aims to emphasize the importance of considering L2 pragmatics and the adoption of a corpus-based approach in the English as a Foreign Language (EFL) classroom by exploring the interlanguage features of learners' disagreements in EFL, and their perceptions of these communicative acts therein. Pragmatics here refers to the linguistic resources for conveying communicative acts and

relational or interpersonal meanings in a language, and the social perceptions underlying interlocutors' interpretation and performance of communicative action (Kasper & Rose, 2001). More specifically, this research attempts to gain insight into learners' L2 pragmatics, so that useful information is provided that can help teachers raise their EFL students' awareness of pragmatic issues in the target language. To this end, disagreements in a corpus of 28 EFL face-to-face conversations of 30 minutes to 1 hour duration each were analysed, and then used in the EFL classroom to examine learners' perceptions of these communicative acts in the target language and generate discussion. In general, a different use of mitigation devices in EFL disagreements was observed in contrast with English native speakers' production of these communicative acts (García, 1989; Kreutel, 2006). Learners therefore showed lack of awareness of the linguistic resources commonly employed for voicing disagreement in the target language. As for their perceptions and discussion of EFL disagreements in the classroom, learners viewed these communicative acts in the target language as adequate and polite at a social level on the whole, which can be said to reflect somehow their L1 pragmatic assumptions on disagreement performance (cf. Cordella, 1996). However, they mostly perceived EFL disagreements as inadequate and impolite at an individual level, thereby evincing pragmatic assumptions typically associated with these instances of communicative action in L1 English (cf. Locher, 2004; Pearson, 1986; Pomeranz, 1984). These findings suggest that a closer look at learners' productions and perceptions of target language behaviour using learner corpora in the classroom can be useful to achieve a better understanding of our students' L2 pragmatics, and help them in their development of target language proficiency.

Gayo, Iria and Luz Rello

Panel: 5. Corpus, estudios contrastivos y traducción

DIFERENCIAS EN EL PÁRAMETRO PRO-DROP ENTRE PORTUGUÉS BRASILEÑO Y ESPAÑOL UTILIZANDO CORPUS COMPARABLES

Tanto el español como el portugués son lenguas pro-drop (Chomsky 1981). No obstante, diversos estudios (Barbosa 2003, 2005) indican que el portugués muestra diferencias respecto a este parámetro entre sus variedades europea y brasileña (de ahora en adelante se utilizará PE para el portugués europeo y PB para su variante brasileña). Esta diferencia radica en que la variedad portuguesa muestra una tendencia más acentuada hacia la substitución del sujeto nulo por formas explícitas (Duarte 1993, 1995). El objetivo de este trabajo es delimitar y describir las naturaleza peculiar del parámetro pro-drop del PB frente a otra lengua que omite el sujeto, el español. Para llevar esto a cabo se ha realizado un estudio basado en la comparación de textos de medicina de dos corpus comparables, uno en español y otro en PB. Las categorías de sujeto explícito, sujeto omitido y ausencia de sujeto (oraciones impersonales), así como las pasivas reflejas (con sujeto explícito o implícito) fueron anotadas manualmente en los corpus por cuatro anotadores diferentes. Los dos corpus están compuestos por textos escritos originalmente en las dos lenguas (no traducciones) pertenecientes a los mismos géneros (género legal y medicina). Cada corpus cuenta con alrededor de 6000 anotaciones manuales. Las categorías tenidas en cuenta han sido las siguientes:

— Sujetos explícitos*

— Sujetos nulos

— Pasivas reflejas*

— Impersonales

* formas nominales o pronominales

Los resultados obtenidos indican que, al igual que entre PB y PE, existen diferencias entre español y PB en lo que a la omisión o del sujeto se refiere. El español se aproxima más a la variante peninsular que a la brasileña, ya que tiende más que este último a la utilización de sujetos nulos (31%) frente a los sujetos explícitos (54%). En la línea de lo que muestran los estudios citados, el PB se inclina preferentemente

hacia el uso del sujeto explícito (70%), siendo el sujeto nulo menos común que en español (19%). De la misma manera, nuestro estudio muestra que también existen diferencias entre ambas lenguas en el uso de la pasiva refleja y las impersonales, ya que ambas categorías son más comunes en el español que en el PB. Finalmente, se muestra una tipología de casos que reflejan la diferencia en la omisión del sujeto en ambas lenguas.

Gil-Salom, Daniela

Panel: 8. Los córpora y la adquisición y enseñanza del lenguaje

LA ADQUISICIÓN DE ALEMÁN COMO LENGUA EXTRANJERA. UNA APORTACIÓN BASADA EN CORPUS DE APRENDICES.

En este trabajo revisamos los estudios relativos a la adquisición de alemán como segunda lengua (L2) y como lengua extranjera (LE), partiendo del trabajo de Clahsen et al. (1983), punto de referencia obligado en cualquier investigación sobre este tema. En primer lugar, describimos los estudios que obtienen las mismas secuencias de adquisición en la sintaxis que Clahsen et al. (1983) en su proyecto ZISA. Estos estudios (Ellis, 1992; Tschirner, 1992; y Boss, 1996; entre otros) afirman que las secuencias de adquisición de las sintaxis son: SVO > SEP > INV > VEND. En segundo lugar, los trabajos de Du Plessis et al. (1987), Boss (2004) y Lund (2004) demuestran que dichas secuencias no siempre se cumplen. Para los primeros, la dificultad en adquirir VEND es superior a la de INV. Para Boss (2004) la adquisición de ambas estructuras puede darse simultáneamente. El estudio de Lund (2004) conduce a esta misma conclusión y no observa ninguna escala implicacional entre las dos estructuras. Revisamos, en tercer lugar, los estudios que han trabajado con córpora de estudiantes de español como lengua materna (L1). Entre ellos, solamente el trabajo de Ehlers (2001) afirma las secuencias del proyecto ZISA, mientras que Grümpel (2004) y Martínez Adrián (2004/05) advierten también una adquisición previa de VEND a INV. Estas dos últimas investigadoras refuerzan más la variación en la producción de las mencionadas estructuras sintácticas en sus últimos trabajos (Martínez Adrian, 2008:34; Grümpel, 2009:146). Para finalizar la revisión bibliográfica incluímos, en cuarto lugar, las investigaciones que analizan tanto la producción sintáctica, como la morfológica. Aunque autores como Jordens (1988) y Vainikaa y Young-Scholten (1994) defienden una relación entre ambos aspectos, otros muchos, como Boss (1998), Tschirner (1999), Diehl et al. (2000), Meerholz-Härle y Tschirner (2001) y Ballestracci (2008) han observado lo contrario. Dada la heterogeneidad de los datos recogidos respecto a las secuencias de adquisición, los distintos córpora (entrevistas, pruebas metalingüísticas y textos escritos), los diferentes niveles de conocimientos previos y el reducido número de sujetos en muchos de los estudios, nos planteamos si dichos resultados pueden aplicarse a nuestro contexto. Para nuestro estudio hemos analizado la interlengua (IL) de 66 estudiantes de la Universidad Politécnica de Valencia (UPV), que aprenden la lengua alemana como segunda o tercera LE. A pesar de estudiar en la misma universidad y responder a un perfil de individuos similar, venimos observando que existen algunas diferencias entre los estudiantes de los distintos centros. Además, al contar con profesores distintos y materiales docentes distintos, los resultados que obtengamos seran más fiables, puesto que no se reducirán a unas condiciones únicas.

Giménez, Pau, Joan Costa, Aina Labèrnia and Àlex Alsina

Panel: 3. Estudios gramaticales basados en córpora

EL PROYECTO DELADI: EVALUACIÓN DEL CONOCIMIENTO Y USO DE LOS PRONOMBRES RELATIVOS EN CATALÁN

El proyecto Deladi (2007-2010, Ministerio de Educación y Ciencia: HUM2007-61916/FILO) pretende evaluar el conocimiento y uso de los pronombres relativos en 26 estudiantes catalanoparlantes de primer curso de la licenciatura de Traducción e Interpretación en la Universidad Pompeu Fabra. El estudio se ha basado en un corpus elaborado a partir de datos que reflejan cuatro estilos de lengua con un grado ascendente de control sobre el propio discurso:

a) Mínimo durante las entrevistas realizadas, en las cuales al cabo de 30 minutos de hablar sobre aspectos cotidianos y en un tono distendido, el entrevistado presta poca atención a cómo dice las cosas.

b) Un poco más alto en las redacciones que realizaron los entrevistados al término de la entrevista, por lo que conlleva de por sí de mínima reflexión el acto de escribir.

c) El más alto en la producción en los ejercicios (uno de traducción español-catalán y otro de rellenar vacíos), ya que el entrevistado tiene que decidir qué forma usa o cómo traduce las frases.

d) Y el máximo en las encuestas sobre gramaticalidad, uso y distribución estilística de los relativos que contestaron, ya que el entrevistado ya no produce sino que directamente se le pide que valore unas cuantas estructuras de relativo.

Una vez recogidos los datos, se realiza un etiquetaje manual de la estructura sintagmática de los antecedentes y las estructuras correferentes. Esto se realiza mediante la creación de botones con el programa Markin para cada tipo de etiqueta que queremos aplicar. También se anotan los metadatos (autor, ejercicio, grado de formalidad, etc.) a mano. A continuación se transforma el archivo de formato Markin a un formato .txt y se aplica el analizador morfosintáctico CatCG (Catalan Constraint Grammar) al corpus. Finalmente, desambiguamos manualmente las categorías morfológicas o sintácticas que nos interesa estudiar y se anotan los rasgos semánticos del antecedente, así como la relación semántica del relativo con la oración principal. Se han completado estas operaciones sobre los ejercicios gramaticales y sobre las redacciones. Mediante la Interfaz de Acceso a Corpus (IAC), desarrollada por la Fundación Barcelona Media y la Universidad Pompeu Fabra, se pueden realizar consultas por forma, lema, función gramatical, estructura del antecedente, relación semántica y demás valores anotados, para poder evaluar el conocimiento y uso de los relativos con relación a factores estilísticos o formales. En esta comunicación vamos a presentar los resultados obtenidos hasta el momento mediante la aplicación de fórmulas estadísticas basadas en el coeficiente de implantación propuesto por M. Amor Montané (2007) en el ejercicio de rellenar huecos. Los resultados provisionales indican que el grado de implantación es alto, hecho por otra parte predecible, puesto que el estudio desarrollado hasta la fecha se ha realizado a partir de los ejercicios de producción totalmente planificada.

Goethals, Patrick

Panel: 5. Corpus, estudios contrastivos y traducción

DEMONSTRATIVE MODIFIERS AND DEFINITE ARTICLES IN TRANSLATION: A CONTRASTIVE PERSPECTIVE.

In this paper I will elaborate a contrastive linguistic analysis of the alternation between the demonstrative modifier (este/ese/aquel problema) and the definite article (el problema) in Spanish and Dutch. The methodology is based on a bidirectional corpus of translated texts (Spanish-Dutch and Dutch-Spanish). Several studies have focused on the semantics of the demonstrative paradigm, in order to distinguish it from the definite article. These studies usually adopted a monolingual or a generic point of view. Instead, very little is known about specific contrastive differences: do both categories relate to each other in a similar way in different languages? The data that come from the bidirectional corpus of translated texts suggest that Dutch and Spanish indeed differ significantly. Concretely, Dutch demonstratives appear to be more broadly used than their Spanish counterparts, and therefore quite often correspond to a definite article in the Spanish source or target text. In the corpus this becomes clear when translational shifts are considered. From a quantitative point of view, the following observations can be made:

1) in the Spanish-Dutch subcorpus, Spanish demonstrative modifiers are rarely translated by a Dutch definite article (23 examples, or 5,7% of the Spanish demonstrative modifiers). Far more frequently, a Dutch demonstrative was newly introduced to translate a Spanish definite article (110 examples, or 21,5% of the Dutch demonstratives).

2) in the Dutch-Spanish subcorpus, the same tendency is found: there are relatively few examples of Dutch definite articles being translated by a Spanish demonstrative modifier (16 examples, or 4% of the Spanish demonstratives), and a relatively high number of cases where a Dutch demonstrative modifier is translated by a Spanish definite article (81 examples, or 17% of the Dutch demonstratives). The fact that the same tendency is found in both subcorpora is important, since it suggests that these translational shifts are not to be seen as a translation universal, but instead as a consequence of a contrastive difference between the two languages. Although the main part of the paper will be dedicated to the methodological implications of the use of bidirectional translational corpora, and to the presentation of the quantitative results of the corpus study, I will also present a qualitative, semantic analysis of some recurrent shifts. In general, there seems to be some evidence that, compared to Spanish, the Dutch demonstrative can be more easily used with an identifying function, instead of the typical reclassifying function of demonstratives. In Spanish, this identifying function would be rather the exclusive domain for the definite article. This semantic analysis might account for shifts such as (1) or (2):

(1)

ES [entonces] no habría sanciones y los gringos pendejos no joderían con la soberanía (Fiesta de Dumas)

NL [dan] zouden er geen sancties zijn en zouden die klotegringo's niet zitten te zeiken over soevereiniteit [= esos gringos pendejos]

(2)

ES - ¿Cómo será? Espero que no sea como las otras. (Medeplichtige)

NL 'Hoe zou ze zijn? Ik hoop niet zoals die anderen.' [= esas otras]

Gómez, Angeles

Panel: 5. Corpus, estudios contrastivos y traducción

CORPUS STUDY BETWEEN THE ENGLISH GERUND AND ITS SPANISH COUNTERPARTS

The previous contrastive studies between the English gerund and its Spanish counterparts present limitations in two specific areas. Firstly, the previous studies do not include all the translation possibilities or counterparts (Alonso García 2003; Izquierdo 2006 and 2008; Losada Durán 1980; Piñeiro and García 2001). In fact, according to our corpus data, in comparison to previous studies, it can be ascertained that the English gerund displays a greater variety of counterparts of a varied nature. In the second place, we have proven that most of the previous studies do not include a cognitive characterization of the English gerund and its counterparts; our work includes a conceptual description of the English gerund and its counterparts. We argue that it is important to include a cognitive description because this description facilitates us to establish a hierarchy between the English gerund and its counterparts based on their coincidences and differences from a cognitive point of view. In this sense, the use of a parallel corpus enables us to check in greater depth the cognitive relationship of the English gerund and its Spanish counterparts. This confirms that a parallel corpus is a suitable tool when carrying out a contrastive analysis. Thanks to the corpus, we have carried out two different studies (the English gerund and the Spanish counterparts) which, in turn, complement each other and confirm part of our hypothesis and also provide interesting results in the field of translation. We have defined the English gerund according to its nominal profile, as an abstract entity, based on cognitive grammar and psycho-mechanical observations (Langacker 2008, and Duffley 2003 and 2006 respectively). The analysis of the counterparts highlights the validity of analysing the English gerund from a nominal profile. In fact, from a conceptual point of view the most frequent counterparts, the infinitive and the substantive share with the English gerund the abstract region's interpretation. According to our corpus data, it can be ascertained that the majority of the most frequent counterparts can be predicted within the Spanish system and show a syntactic and semantic independence in opposition to the English gerund. As the analysis progresses, we observe the frequency of less predictable translations which put the role

orthonymy into play. The concept of orthonymy designates the most habitual, natural and authentic way of expressing yourself in a language. In these cases, in general, it is corroborated that the Spanish translation "distances itself" from the linguistic system of the source language in favour of a more authentic translation of the target language. First, we will provide the cognitive characterization of the English gerund. To follow, we will present the counterparts in terms of their cognitive coincidences and differences in relation to English gerund. And, finally we will provide a translation approach by which a particular Spanish counterpart can be explained.

Gonzálvez-García, Francisco

Panel: 5. Corpus, estudios contrastivos y traducción

"THE GRAMMAR-DISCOURSE INTERFACE REVISITED WITHIN CONTRASTIVE CONSTRUCTION GRAMMAR: THE CASE OF FOCUS CONSTRUCTIONS IN ENGLISH AND SPANISH"

This paper argues for a bottom-up, usage-based constructionist account à la Goldberg (1995, 2006) of Spanish contrastive focus configurations of the type exemplified in (1)-(2).

(1) Oh Sir Salman, you CERTAINLY know how to charm a gal (The Sunday Times, June 1 2008) (http://www.timesonline.co.uk/tol/comment/columnists/article4040060.ece)

(2) TÚ sí que sabes, Woody (http://cine.linkara.com/pelicula/annie_hall/critica/150360/tu_si_que_sabes_woody/)

Specifically, compelling evidence is adduced for the existence of a number of non-trivial analogies regarding (i) the core meaning of the constructions, (ii) the semantico-pragmatic profile of the Element in Focus (henceforth EIF), (iii) their newness orientation and (iv) their (positive/negative) interpersonal flavour and (v) their thematic and cohesive flexibility, inter alia, which enable us to treat the constructions in (1)-(4) as forming a family (or constellation) of constructions. First, it is argued that the core constructional meaning of focus constructions is to provide the identification by the speaker/writer of an entity (person or thing) (i.e. the EIF = identified at a particular stage in discourse as the Focus of Attention) that is connected with an open proposition that may be equational (as in clefts) or characterizing (as in the other constructions). Furthermore, the constructions under scrutiny here appear to move along a cline (or, alternatively, a path) of referential > non-referential functions (see further Dasher 1995). Second, The EIF is more likely than not referential and specific, which means that subject expressions in idiom-chunks or non-specific expressions are ruled out in the slot in these constructions, as shown in (5). Third, following Zimmerman (2007: 158), it is argued that 'newness' "must take into account discourse-pragmatic notions like hearer expectation or discourse expectability of the focused content in a given discourse situation. The less expected a given content is judged to be for the hearer, relative to the Common Ground, the more likely a speaker is to mark this content by means of special grammatical devices, giving rise to emphasis." However, the examination of the constructions at hand here shows that this should be best regarded as a tendency rather than as a cut-and-dried generalization. Finally, regarding their interpersonal nature, contrastive focus constructions in general and clefts in particular convey a positive or negative stance by subject/speaker towards the content of the proposition. Therefore, these constructions encode a higher degree of subjectivity (i.e. emotional intensity) and convey a slightly more accusatory or a slightly more laudatory tone than their non-cleft/non-focus counterparts (Perzanowski & Gurney 1997: 221-222). Thus, in a (6), the pressure exerted by Carter is considered to be the major driving force to free Nicaragua from censorship. This semantico-pragmatic facet of the constructions under scrutiny here can be grounded on the notion of subjectivity, viz. "the way in which natural languages, in their structure and normal manner of operation, provide for the locutionary agent's expression of himself and his own attitudes and beliefs." (cf. Lyons 1982: 102; Scheibman 2002). Finally, regarding thematic and cohesive flexibility, given that contrastive focus constructions can be used to explicitly signal a contrast, alternative, or correction with respect to a previous stretch of discourse, they qualify as cohesion-building devices. A case in point is example (7), where the writer makes recourse to restatement based on a play on words to convey his/her negative stance on the supporters of the Spanish socialist Party.

Goutsos, Dionysis, Constantin Potagas, Dimitris Kasselimis, Maria Varkanitsa and Ioannis Evdokimidis

Panel: 1. Diseño, compilación y tipos de córpora

THE CORPUS OF GREEK APHASIC SPEECH: DESIGN AND COMPILATION

The study of aphasia in Greek lacks large-scale empirical findings, mainly because of the theoretical orientation of the field. Computer language corpora can usefully fill this gap and give a new perspective to the study of the speech of Greek aphasic patients. The paper's goals are to present the design and compilation of the Corpus of Greek Aphasic Speech (CGAS), a new resource for the study of aphasia in Greek, and to discuss its possible applications. The aims and design of the corpus and the methods followed for its compilation are presented. A pilot corpus was first created, including data from 20 patients, treated between 2006 and 2008. Two type texts from each patient's spoken output have been included in the corpus, namely spontaneous speech and picture description (12.663 words, in total, of which 10.332 belong to patients' talk). On the basis of the pilot corpus, a classification of paraphasias or speech errors has been attempted and the frequency and type of each category has been studied. The Corpus of Greek Aphasic Speech is envisaged to include data from 114 patients, that is 228 texts of 50.000 words in total (of which 41.000 spoken by patients). In conclusion, it is argued that the exploitation of specialized computer corpora can have important advantages for the study of aphasia and can usefully complement current research on aphasia in Greek, both quantitatively and qualitatively. Among the most important consequences of using corpora in aphasia research is the view of speech errors as the product of situated language use by specific speakers rather than as isolated examples of lack of competence.

Gregori-Signes, Carmen

Panel: 2. Discurso, análisis literario y corpus

COMMUNITY DIGITAL STORIES: A CORPUS ANALYSIS

Digital Storytelling is genre which is rapidly expanding in many different fields including education, socio-cultural studies, turism and marketing, to mention but some. However despite the variety of digital storytelling "little has been written on digital storytelling, outside the occasional "how-to" guides by practitioners" (Hartley and McWilliam 2009:5). This article seeks to make a contribution in the analysis of community stories to check whether they could be classified as examples of socio-political digital storytelling. Socio-political digital storytelling is here defined as type of digital story which may potentially become a powerful tool that may help bring up and out issues that may concern and affect democracy (Couldry 2008) and social welfare. For the analysis of these community stories I draw upon the principles of critical discourse analysis- this being understood as an approach rather than a method- combined with corpus linguistic methodology; and on the principles of sociopragmatics (Leech 1983: 10) since I believe in the importance not only of studying communication within its sociocultural context, but also in the need to find out the different sociopragmatic rules that may apply when denouncing a situation which affects or affected the author's life in the past (cf. Gregori 2010). The stories analysed in this article have been obtained from the website Australian Centre for the Moving Image (ACMI) and have been transcribed and analysed drawing upon two different corpora: a) the content of a total of 10 websites that admit using digital stories with social purposes; b) a detailed analysis of the topics or semantic macrostructures and of the local meanings (van Dijk 2001:101) of each 25 stories. Due to space restrictions, the analysis here focuses on the study of community stories by looking mainly at the textual structure of the stories and of the web pages, thus paying attention to: a) the topics of the texts; b) the lexical choice or vocabulary in the stories. The hypotheses operating in the analysis can be stated as follows: a) whilst it is probable that each story displays its own idiosyncracies, the results of the analysis should at least shed some light on the factors that may be of interest for the members of a community; b) that although the participants may not all fit the same pattern regarding age, time,

motivation to write the story, and physical, intellectual, linguistic, social, cultural and emotional development, among others, a corpus analysis of their content should show a relation of topics/ vocabulary of the social representations (van Dijk 2001:113), the knowledge, attitudes, ideologies, norms and values of the social order which they abide. If that were the case, not only would the hypotheses be confirmed; but, secondly, this would prove that corpus analysis may be considered as a valid tool to find out more about the nature of different types of digital stories.

Grochocka, Marta

Panel: 4. Lexicología y lexicografía basadas en córpora

NONCE FORMATIONS AS INDICATORS OF PRODUCTIVE WORD-FORMATION PROCESSES IN ENGLISH

Coinage, borrowing and word formation are the three major methods of extending the lexicon, with the last one being the most productive. In other words, the highest proportion of neologisms come into existence as a result of word-formation processes in which already existing elements of a language are manipulated in some creative way. Every neologism begins its lifecycle as a nonce formation which is created as a consequence of satisfying a particular communicative need arising on a particular occasion. To begin with, it is crucial to make a clear distinction between nonce formations and neologisms as there is considerable terminological confusion in the literature. Another problem is that nonce formations themselves may be perceived in two opposing ways, i.e. as ad hoc, context-dependent and non-lexicalizable deviations from word-formation rules (Hohenhaus 1998), or quite the contrary, as formations which are regular, structurally transparent, productively coined and hence predictable (Štekauer 2002). The latter viewpoint is adopted in the present study. Moreover, being indicative of productive word-formation rules, nonce formations are believed to be worthy of study, although they are often transient creations with little chance of becoming institutionalised. Additionally, various types of nonce formations are discussed, with context-dependent naming units and neologistic wordplay as the prime focus of interest. A web-based application called NeoDet has been developed for the purpose of compiling a study corpus of journalistic texts and extracting neologism candidates from the corpus, among which a host of nonce formations and wordplay units can be found. The three-million-word corpus consists of articles and blogs from the most widely read British newspapers and tabloids (i.e. The Daily Telegraph, The Times, The Guardian, The Sun, and The Daily Mail) published between 1st January and 31st December 2009. The neologism candidate detection procedure is based on the exclusion principle, with the exclusion sources including a few online dictionaries (i.e. OALD7, MW11, MEDAL2, CH11, CALD3, LDOCE5, Google Dictionary and dictionary.com), four slang dictionaries, the British National Corpus, as well as a wordlist of proper names and geographical names. A lexical item is regarded as a neologism candidate only when it is absent from all the exclusion sources. Once a nonce formation coined by means of affixation has been discovered, the NeoDet search engine is used in order to establish the degree of productivity exhibited by a given prefix or suffix. In this way, studying nonce formations makes it possible to uncover English productive affixes and draw conclusions concerning their meanings. Furthermore, the study sheds light on certain strategies adopted by journalists with the aim of attracting public attention. All in all, new naming units are coined not only to compensate for the denotational deficiency of a language, but also with the purpose of being eye- and ear-catching, witty, amusing and memorable.

Guerrero Triviño, José María, Rafael Rafael Martínez Tomás, Mª Carmen Mª Carmen Díaz Mardomingo and Herminia Peraita Adrados

Panel: 9. Usos específicos de la Lingüística de Corpus

MODELO DE RED BAYESIANA BASADO EN UN CORPUS LINGÜÍSTICO DE DEFINICIONES CATEGORIALES APLICADO AL DIAGNÓSTICO DEL DETERIORO SEMÁNTICO COMPATIBLE CON DEMENCIA TIPO ALZHEIMER

Las técnicas de Inteligencia Artificial, como las Redes Bayesianas, pueden contribuir al diagnóstico de la enfermedad de Alzheimer (EA), por ello hemos empleado un modelo de Red Bayesiana, basado en el Corpus Lingüístico de definiciones orales (Peraita y Grasso, 2009) http://www.uned.es/investigacion-corpuslinguistico/. Este Corpus supone un instrumento metodológico de primer orden para el estudio de enfermedades que cursan con deterioro semántico. La Red presenta un modelo causal basado en el Corpus de definiciones de categorías semánticas -seres vivos y seres no vivos-. En la EA se produce un deterioro semántico diferencial entre ambos tipos de dominios categoriales. Generalmente hay una mayor afectación del conocimiento de los seres vivos mientras que el de los seres no vivos está más conservado, aunque también hay evidencia del patrón opuesto (revisión de Capitani, Laiacona, Mahon y Caramazza, 2003). Las Redes Bayesianas constan de dos componentes: la estructura y los parámetros. La estructura -parte cualitativa- define las relaciones causales, funcionales e informativas, identificadas en el dominio. Los parámetros son las probabilidades condicionales y utilidades, y constituyen la parte cuantitativa que expresa la fuerza de las relaciones probabilistas siendo representadas por probabilidades condicionales. Las relaciones causales entre variables suelen acompañarse de un factor de incertidumbre, que se puede expresar a través de la fuerza de la relación. Las Redes Bayesianas son extremadamente útiles en la respuesta ante nuevos casos, y existen técnicas de Aprendizaje Automático que permiten descubrir nuevas relaciones entre variables o nuevas probabilidades condicionales según aparecen nuevos casos. Nos proporciona un diagnóstico y una gran capacidad analítica, permitiendo expresar matemáticamente la posible influencia de nuevas variables en el diagnóstico. Las definiciones que forman el Corpus, fueron producidas por personas mayores sanas y enfermos de Alzheimer de España y Argentina. Se solicitó a los sujetos que definieran seis categorías, tres de seres vivos y tres de no vivos, las cuales fueron grabadas y transcritas para su análisis cuantitativo (frecuencias de producción de rasgos, para cada categoría, etc.) y cualitativo (diferentes tipos de rasgos según modelo de Peraita, Elosúa y Linares, 1992). Este análisis proporciona las evidencias para la Red Bayesiana. En el modelo causal representamos que la EA es causa de un déficit léxico-semántico-conceptual y la Red Bayesiana inferirá la probabilidad de padecer la EA, a partir del grado de dicho déficit. Este modelo se basa en un razonamiento abductivo, se parte del deterioro semántico y se busca la probabilidad de que ese deterioro explique el padecer EA. Se aborda, la lógica que subyace al análisis de rasgos propuestos, según el modelo de Peraita et al. (1992), en la línea de otros trabajos (Cree y McRae, 2003; McRae et al. 2005; Peraita y Moreno, 2006). Los objetivos de este trabajo son: a) empleo de un modelo basado en Redes Bayesianas para el diagnóstico del deterioro semántico; uso del aprendizaje automático del modelo cuantitativo, a partir de una base de casos y de estudios epidemiológicos; c) análisis de sensibilidad de evidencias; d) análisis de sensibilidad de los parámetros; e) interfaz de usuario en Web. Se presenta el modelo y las decisiones que se han tomado para llegar a él.

Gutiérrez, Camino

Panel: 1. Diseño, compilación y tipos de córpora

FROM CATALOGUE TO CORPUS IN DTS: TRANSLATED AND CENSORED CINEMA UNDER FRANCO (TRACECI 1951-1962)

One of the main proposals of Descriptive Translation Studies (DTS) is that, in order to obtain relevant results, we need to carry out a systematic study of those original and translated texts that, far from being chosen at random, have been carefully selected following certain well defined criteria. Textual selection should, therefore, be considered as one of the key stages of the research. This presentation aims at highlighting the role of TRACE* Catalogues as an essential tool in textual selection, by describing the transition from Catalogue to Corpus in the study of translated and censored cinema under Franco during the 50s and 60s, which is part of the research that has been carried out by the TRACE (translation and censorship) project for more than ten years. In the current TRACE Catalogues of translated and censored narrative, theatre, poetry and audiovisual (cinema and TV) texts, "each individual target text is accounted for in a single record, that contains both contextual and pre textual information related to that target text. This is what makes TRACE database a potential matrix for the selection of corpora (Merino 2001), and why each catalogue can be defined as zero-corpus" (Merino 2005). Their compilation has been done by systematically feeding them with the information gathered from both

censorship archives and other sources of information. The TRACEci 1951-1962 Catalogue currently holds around 3,500 entries, with useful pre/contextual information about the films that were translated (mainly dubbed) from English into Spanish, censored, and shown in the Spanish screens from 1951 to 1962. From the analysis of the information recorded in the Catalogue, certain sets/chains of source and target texts can be identified as prototypical examples depending on the purpose of the analysis, that is, depending on the different translation and censorship phenomena worth studying: for example, the effect of official and/or religious censorship, the translation and censorship of different genres, different types of films (the so-called "commercial films" or "films of special interest"), etc. Our presentation will show the way the TRACEci 1951-1962 Catalogue has been compiled and the way it has been analysed in order to identify certain texts which will be part of the TRACE parallel corpus and will, therefore, become the objects of close study.

Hedeland, Hanna

Panel: 1. Diseño, compilación y tipos de córpora

INTERACTION OF TECHNOLOGY AND METHODOLOGY IN BUILDING AND SHARING AN ANNOTATED LEARNER CORPUS OF SPOKEN GERMAN

This paper discusses the technological and methodological challenges in creating and sharing HAMATAC, the Hamburg Map Task Corpus. In the first part of the paper, I will introduce the HAMATAC corpus, which consists of 24 recordings of advanced German learners solving a map task (Brinckmann et al. 2008) in pairs. It also includes metadata on all speakers' language biographies. The first corpus version, consisting of original recordings, orthographic transcriptions and metadata, is publicly available. Future versions will include annotations describing various linguistic levels and phenomena – the more subjective in nature, the more interesting from a methodological perspective. Currently we are annotating disfluencies, one example of such subjective phenomena, using an annotation scheme with necessarily interpretative categories. The corpus presentation will also include an overview of EXMARaLDA, which was used to create the HAMATAC corpus. The EXMARaLDA system consists of data models, formats and tools for transcribing, annotating, managing and analysing spoken language corpora with help of three software components: The Partitur-Editor, a tool for transcription and multi-level annotation of digital audio or video recordings, the Corpus Manager, a tool for compiling recordings and transcriptions into a corpus and managing corpus metadata, and EXAKT, a tool for carrying out queries and analyses. I will demonstrate how these components are used for corpus building and to analyse corpus data. I will also describe how the entire set of digital data can be transformed into formats independent of these tools and shared with others via a website. In the second part of the paper I will use HAMATAC to discuss different solutions to some recurrent methodological issues in corpus building and sharing and show how technological and methodological aspects can be said to interact.

- One of the most fundamental questions arises from the non-trivial problems inherent in transcribing spoken language in general and learner language in particular – how do we represent the non-standard characteristics of the data?

- Do the possibilities resulting from technological advances – extensive querying of linguistic data or integrated audio or video in a transcript – affect choices regarding the visual representation?

- How can we ensure comparability with other digital corpora, yet without the restriction of shared transcription conventions?

- How do we implement and apply annotation schemes with various layers, different types of annotations, possibly overlapping each other across and within layers?

- How can we assess transcription and annotation quality when our annotation categories, as in the case of disfluencies, are inherently interpretative?

- How do we establish guidelines clear enough to allow for intersubjectivity and thus for each manual annotation task to be replicable?

- And how do we ensure our corpus project results in a sustainable language resource?

In this sense, I will argue that the interaction with technological aspects plays an important role in further developing the methodology of linguistic corpus building and sharing.

Illamola, Cristina

Panel: 6. Corpus y variación lingüística

LA INFLUENCIA DE LA L1 EN EL USO DE LA CONSTRUCCIÓN "IR A + INFINITIVO" CON VALOR PROSPECTIVO EN LAS ZONAS BILINGÜES

En diversas zonas de Hispanoamérica, la sustitución del futuro sintético (FS) en -ré (cantaré, lloverá) por la construcción "Ir a + Infinitivo" (voy a cantar, va a llover) resulta cada vez más evidente. Si bien en las zonas peninsulares monolingües esta sustitución también se percibe, no es así en las zonas bilingües en las que el español está en contacto con el catalán (1). En esta ocasión pretendemos verificar si los hablantes con el catalán como L1 emplean en menor medida la construcción "Ir a + Infinitivo" para expresar valores temporales prospectivos. Para ello, nos valdremos del corpus RESOL; un corpus de datos orales compuesto por entrevistas semidirigidas realizadas a niños en 6º de primaria, y nuevamente en 1º de la ESO, de escuelas de Mataró (Barcelona). Tras el análisis del corpus, los datos revelan que, efectivamente, en zona bilingüe, los informantes con el catalán como L1 tienden a emplear el FS en mayor medida que la construcción perifrástica. En cambio, los hablantes cuya L1 es el español realizan un uso mayor de la perífrasis. No obstante, este uso no es tan profuso como en el resto de zonas monolingües peninsulares. En definitiva, el hecho de tener el catalán como L1 parece ser el factor que condiciona la proliferación de la construcción Ir a + Infinitivo en el español hablado en Cataluña. Concretamente, el paradigma verbal particular de catalán y el hecho de que anar a + Infinitivo no haya gramaticalizado los mismos valores que el español confieren a Ir a + Infinitivo un uso particular en las zonas bilingües.

Iria Romay

Panel: 6. Corpus y variación lingüística

A PRELIMINARY STUDY OF NEUTRAL MOTION VERBS IN LOB AND FLOB

The semantic domain of motion and space has been exhaustively studied in the last decades, being considered a cognitive universal, together with colour terms or terms referring to family members, among others. Research in the particular field of motion is mainly based on Talmy's (1991, 2000, 2007) typological classification of languages into Satellite-framed (S-languages) and Verb-framed (V-languages). The difference here lies in the lexicalization of the path of motion. If one language codifies or 'frames' a path within the verb (e.g. Spanish María cruzó el parque), then it is a 'verb-framed' language, whereas if it codifies path through satellites (e.g. English Mary walked across (the park)), it is referred to as being 'satellite-framed'. Thus, motion events in V-languages are typically expressed by the combination of a path verb and a subordinate adverbial of manner, in contrast with S-languages, which express them by means of a manner-motion verb and a path satellite. In keeping with the abovementioned typological differences, V-language users tend to encode fewer path segments than S-language users in both speech and written language. Moreover, in S-languages, path information is expressed in a more compact way than in V-languages. Therefore, there seems to be general agreement on the supremacy of English (S-language) over Spanish (V-language) in the expression of motion events, since English makes use of more fine-grained distinctions, especially if we consider motion verbs which also imply manner meanings. These verbs are used much more widely than their Spanish counterparts and can occur in a wider number of contexts. Thus, apparently, and due to lexicalization patterns, there exist remarkable differences between the two languages in what concerns the variety of verbs

expressing manner of motion. The pilot research presented in this paper is part of a larger project whose aim is to provide a contrastive analysis of the development of verbs of manner of motion in English and Spanish as represented in different corpora. There are indications (see, for instance, Martínez Vázquez 2001) that usage in the field of motion may be undergoing change, particularly in Spanish, as a result of contact with or borrowing from English, but also in English itself. In this preliminary study, however, the focus will only be on the English field of motion along the diachronic dimension. For this purpose, three neutral English run verbs (walk, run, and jump) that express manner of motion have been taken into consideration by comparing two sub-periods of Present-day British English (the 1960s and the 1990s) as represented in the LOB and the FLOB corpora respectively. These three verbs have been selected on the basis of their frequency and also because they are generally used in sentences which provide movement information through the verb itself or through other parts of the sentence (the information provided does not only refer to the subject entity but also to manner, path and ground). Therefore, run verbs can be considered one of the core elements in spatial semantics when expressing change of location.

Ivanova, Anna

Panel: 2. Discurso, análisis literario y corpus

PRESIDENTIAL SPEECH IN 140 SYMBOLS: A CROSS-CULTURAL ANALYSIS OF TWITTER USE BY BARACK OBAMA&DMITRIY MEDVEDEV.

The present study is a continuation of a pilot project on the use of Twitter by Barack Obama. As it was proposed elsewhere (Ivanova 2011: in press), a cross-cultural comparative analysis was necessary to get a complete understanding of political talk online as a phenomenon of the 21st century. For this purpose we collected a corpus of Twitter messages (English version) posted by Russian President Dmitry Medvedev who opened his Twitter account during an official visit to the USA in June 2010. Thus, updated corpus comprises 831 tweets posted by Russian and American Presidents during the period June-January 2010-2011.

The analysis shows:

1. Twitter use does not coincide with presidents' work weeks;

2. a slight decrease in Twitter use by Russian leader, while his American colleague sticks to a steady rhythm. Mean for tweets per month: Obama 64, Medvedev 40, i.e. Obama posted 1.6 more tweets;

3. 0.68 of all Obama's messages contain external links; while Medvedev's Twitter has only 0.27 of them (0.61 - are president's photos);

4. low lexical density of corpora: 0.19 (Obama), 0.31 (Medvedev);

5. mean for characters:

a. Barack Obama: 120 (range: 41-140); mode=139; StDev=21,63;

b. Dmitry Medvedev: 116 (range: 16-140); mode=140; StDev=24,86;

6. Gunning-Fog Index: 14.8 (Obama), 16.8 (Medvedev);

7. high usage of "we" (N=128), "watch" (N=97) and "live" (N=95) in American corpus; and of "we" (N=63), "Russia" (N=30) and "today" (N=29) in Russian one;

8. the most frequent collocates of node WE within the span 4:4 are:

a. in Obama's Twitter: WE 128 <can 22, our 13, win 13, need 12, move 10, forward 10, you 10, america 9, your 8, fight 7>;

b. in Medvedev's Twitter: WE 63 <have 15, 9 need, discussed 6, agreed 4, issues 4, summit 4, working 4, energy 3, global 3, russia 3>

Thus, we conclude that:

1. Twitter use by both presidents presents a monodirectional interaction channel where Twitter platform is used as an advertisement tool to give an additional promotion to presidents and their cabinets' actions;

2. Nearly maximum use of available symbols proves an extensive use of Twitter by both presidents;

3. According to readability index both Twitter corpora are classified as technical documents, i.e. their target audience is expected to have a university degree;

4. The lexical component of both Twitter corpora is restricted to the professional side of presidents' political actions and excludes any other type of information, i.e. there are no chunks containing other type of vocabulary which we then consider as lexically even distributed.

This continuation of a previous study proves Twitter to be a useful online social platform which serves as an additional promotion tool in the domain of political communication. Its language component does not go beyond political vocabulary which is then seen as lexically limited. Thus, we see that new technologies are used to tell basically the same "old" story but in modern and fashionable frame.

Izquierdo Alegría, Dámaso and Ramón González Ruiz

Panel: 2. Discurso, análisis literario y corpus

CORPUS PARALELOS Y ANÁLISIS DEL DISCURSO: PROPUESTAS DE EXPLOTACIÓN A PARTIR DEL ESTUDIO DE UN MECANISMO COHESIVO

Un corpus paralelo es un tipo de corpus que, en palabras de McEnery y Xiao (2007: 2), contiene "source texts and their translations". La existencia de corpus paralelos ha supuesto un importante cambio tanto en la praxis traductora como en la investigación traductológica. En efecto, la principal obra multilingüe de consulta de la que tradicionalmente disponía el traductor era el diccionario, herramienta que describe la lengua en tanto que sistema. La irrupción de los corpus paralelos supone un cambio de paradigma en el ámbito de la traducción, pues estos corpus muestran directamente la lengua en uso en textos concretos a partir de una recopilación de traducciones preexistentes. Por lo tanto, esta situación constituiría un síntoma del asentamiento de los postulados de la Lingüística del Texto en la práctica traductora. Teniendo en cuenta esta base metodológica, no es de extrañar que el Análisis del Discurso pueda sacar gran provecho de los corpus paralelos, pese a que esta herramienta, como manifiesta Baker (2006: 45), no haya sido creada ad hoc para dicha disciplina. Por ello, la presente comunicación trata de presentar las posibilidades que ofrece para la investigación en el Análisis del Discurso la explotación de una herramienta aún poco conocida más allá de la Lingüística de Corpus y la Traducción: los corpus paralelos. Estas posibilidades se ilustran con varias propuestas de uso en torno al estudio de la anáfora conceptual, dado que sus propiedades discursivas la convierten en un elemento lingüístico especialmente propicio para mostrar las aplicaciones de esta herramienta al Análisis del Discurso. Así pues, las anáforas conceptuales son nominalizaciones que encapsulan fragmentos previos de un texto (función compresora), a la par que tienen la capacidad de introducir nuevas interpretaciones respecto a su antecedente, en función del significado y las connotaciones que transmita la anáfora conceptual escogida (función expansiva). Existen otros modos de hacer referencia a segmentos previos, como las proformas gramaticales y las proformas léxicas, pero que, a diferencia de las anáforas conceptuales, cuentan con un potencial expansivo muy limitado o inexistente. Algunos estudios, sin el amparo de corpus paralelos, han intuido que cada lengua parece mostrar una preferencia hacia uno u otro mecanismo y que su comportamiento presenta ciertas diferencias (Álvarez-de-Mon y Rego 2001, Descombes y Jespersen 1992, Moirand 1973, Peña Martínez 2006, Schmid 2000, entre otros). De este modo, el corpus paralelo abre la puerta a la realización de estudios contrastivos en los que se detecten de manera sistemática estas diferencias en el uso de anáforos a través de búsquedas relativamente sencillas. No obstante, el papel de los corpus paralelos en el Análisis del Discurso trasciende la

perspectiva contrastiva, que es la más frecuente, y cuenta con interesantes aplicaciones en estudios discursivos monolingües: efectivamente, el contraste con otras lenguas facilita la detección de los rasgos diferenciales de estos mecanismos cohesivos en un idioma concreto, sus empleos más habituales y los efectos de sentido que aportan, como trataremos de mostrar en esta comunicación.

Ji, Meng

Panel: 6. Corpus y variación lingüística

A CORPUS-BASED STUDY OF DIACHRONIC REGISTER VARIATION IN MODERN CHINESE

This paper sets out to investigate diachronic register variation in modern Chinese through a corpus-based comparative study of two large-scale monolingual corpora of modern Chinese, i.e. the Lancaster Corpus of Modern Chinese (LCMC) (1990s) and the UCLA Corpus of Modern Chinese (early 2000s). The study of register variation came to prominence in the 1990s with the advent of language corpora and the technical advancement of natural language processing tools. Earlier attempts were made at uncovering the patterns underlying register variation. The patterns thus identified might help establish a multidimensional framework for cross-cultural and cross-linguistic analysis (Biber, 1995). The validity and wider applicability of the model was tested with four orthographically different linguistic systems which were English, Nukulaelae Tuvaluan, Korean, and Somali. It is however argued in this paper that the representativeness of the model thus built requires further verification with language data collected from orthographically similar but socio-culturally different linguistic systems such as Korean and Chinese. That is because the development of modern written registers in these two languages, despite their many shared textual and discourse conventions, may have well followed distinctive patterns of evolution as a result of the different cross-cultural contacts with the West that they were exposed to. Therefore, in this paper, we aim to explore the particular patterns of register variation in modern Chinese within the multidimensional framework of linguistic analysis proposed in Biber (1998). The innovative of relevant corpus data and methods proved essential in the discovery of novel textual and linguistic events bearing on the changing nature of written genres in modern Chinese as documented in the two larges-scale comparable corpora under investigation.

Judith Laso, Natalia, Elisabet Comelles, Isabel Verdaguer

Panel: 8. Los córpora y la adquisición y enseñanza del lenguaje

USING A CORPUS-BASED CLAUSE PATTERN DATABASE IN THE ENGLISH GRAMMAR CLASSROOM

The use of corpus-based tools has proven to be useful for the teaching and learning of a foreign language (Aston 2001, Granger 2003, Sinclair 2004, Conrad 2005, Granger & Meunier 2008, Aijmer 2009, Bennet 2010) as it allows both the linguist and the learner not only to become aware of the complexity of language but consider utterances in a real context as well. Likewise corpus linguistics has stressed the systematic interconnections between lexical items and their linguistic environment. It has empirically shown that native speakers tend to make use of recurrent strings of words, and has greatly contributed to the identification of units of meaning, which would have been hardly detected without the assistance of corpus-based methods. Most corpus-based studies conducted up to now deal with an empirical description of language, yet there are very few studies exploring the benefits of following this approach for language teaching (Conrad 2005, Laso & Giménez 2007 & 2008, Aijmer 2009, Bennet 2010). Although these benefits would seem consistent with language learning theory, little research on the effectiveness of using corpus-based materials in the EFL classroom has been carried out so far. As part of a teaching innovation project devoted to the creation of teaching materials, the GReLiC group at the University of Barcelona has recently developed the Clause Pattern Database (CPDB), which gives account of the valency patterns performed by a selection of 45 prototypical verbs. This corpus-based tool is also supplemented with tree diagrams, created with the assistance of the Charniak parser (Charniak and Johnson 2005) and the PhpSyntax Tree, illustrative of each example in the database. This paper aims at

illustrating the various applications of the CPDB for the teaching and learning of verb subcategorisation requirements. To this end, a continuous assessment task, especially designed for the undergraduate course Descriptive Grammar of English, will be presented. The task was conducted in the 3 groups of third-year students, of approximately 50 students each. In the task students were asked to: a) complete the CPDB with real examples of language (excerpted from texts of their choice) by providing their valency and clause pattern; b) provide a tree analysis of each sentence. Once the task was completed, they were also asked to answer an online questionnaire so as to assess their satisfaction towards the newly designed database and corpus-based activity and explore how corpus linguistics can contribute to language acquisition in formal tuition contexts.

Juncal, Lourdes

Panel: 5. Corpus, estudios contrastivos y traducción

A CONTRASTIVE STUDY OF ADVERBS OF CERTAINTY AS DISCOURSE MARKERS IN SPOKEN ENGLISH AND SPANISH

The present paper will focus on the adverbs certainly, definitely, obviously, and absolutely in British English, and on their equivalents in Castilian Spanish, which I will divide into two groups: 1) their literal equivalents (ciertamente, definitivamente, obviamente and absolutamente) and 2) their equivalents in use (por supuesto, naturalmente, sin duda, claro, desde luego, cierto, etc.). All these adverbs of certainty (Martín Zorraquino & Portolés, 1999; Vandenbergen & Aijmer, 2007) will be analyzed in this presentation as discourse markers which are indexically linked to epistemic modality. The function of these adverbs as discourse markers, working as a whole sentence in conversation, has not been extensively analyzed. The aim of this study is to analyze the speaker´s reactive intervention (Martín Zorraquino & Portolés, 1999) when these markers occur as a whole sentence in turns of talk in order to determine conversational strategies (agreement, indirectness, fluency, interruption, empathetic use, power, solidarity, etc). In addition, I will show their differences and similarities in use and frequency in English and Spanish. By means of the Wordsmith Tools programme I will be able to compile wordlists, frequencies, and concordances in order to analyze grammatical features such as the position of the marker with respect to the discursive member where it occurs. Furthermore, I will examine contextual features to show which markers are used in formal and non-formal registers, as well as gender and age differences in usage. This study will utilize samples taken from two corpora: the Integrated Reference Corpora for Spoken Romance Languages (C-ORAL-ROM) for Spanish, and the London Lund Corpus of Spoken English (LLC) for English. Bearing in mind that these two corpora vary in their quantity of words, I will apply Bibber´s procedure (1988) calculating the frequency of occurrences per million words in order to guarantee a comparable analysis.

Karakoc,Taner

Panel: 5. Corpus, estudios contrastivos y traducción

CORPUS OF TURKISH CULTURE-SPECIFIC ITEMS AS REPRESENTATIVES THROUGH TRANSLATION IN ISTANBUL 2010 EUROPEAN CAPITAL OF CULTURE ACTIVITIES

The paper aims to investigate the function of the corpus of Turkish Cultural Items as representatives of Turkish culture through translations produced during the activities organized within the scope of Istanbul 2010 European Capital of Culture Project. The monthly bilingual (Turkish – English) events bulletins as published online or in a booklet format serve as a means of resource of information for the corpus on the cultural activities held in the project highlighting conferences, concerts, documentary screenings, exhibitions, workshops, drama, nobel ceremonies of Sema, drama, performances etc. Such cultural items, or “culturemes” that make up the corpus convey invaluable information through translation about Turkish culture for foreign viewers. Among such culture-specific terms are cultural items related to music, food, local arts& crafts, traditions, dance, drama, religion, religious ceremonies etc. The study describes the methods of translation (modulation, adaptation, transposition,

explicitation, omission, amplification, compensation, etc) implemented based on the texts appeared in such bulletins, which make up the corpus of the analysis. The study also provides a multifaceted analysis with references to paradigms in Translation Studies such as equivalence, descriptions, purposes, uncertainty and above all, cultural translation (Anthony Pym, Exploring Translation Theories, 2010).

Keshabyan, Irina

Panel: 5. Corpus, estudios contrastivos y traducción

A CONTRASTIVE STRUCTURAL ANALYSIS OF SHAKESPEARE'S HAMLET VERSUS SUMAROKOV'S GAMLET: A CORPUS-BASED APPROACH

The main aim of this paper is to look at the structural (dis)similarities of two specific texts in the genre of drama -The Fourth Folio Edition of The Tragedy of Hamlet Prince of Denmark (1685) by Shakespeare and the English translation of Gamlet (1787) [1748] by the Russian playwright Sumarokov, translated from Russian by Richard Fortune in 1970. The main area of research of this investigation is the study of text by means of corpus-based techniques -in other words, by means of a computational and quantitative analysis. For ease of reference, The Fourth Folio Edition of Shakespeare's Hamlet (1685) will be referred to as Hamlet or SH. The Russian text will be referred to as SG-R, whilst the English translation will be referred to as Gamlet or SG. The investigation is based on the electronic collection of these texts, that is, on the computerised texts. The method I use to analyse Hamlet and Gamlet does not dwell on the standpoints of various forms of historical, philosophical, language-based, etc. approaches which are available at present. So, what I do is focus on the formal aspects of the plays that could be easily located, extracted, computerized, quantified and, at the same time, could contribute towards identifying Shakespeare and Sumarokov's intentions, particularly with regard to the structural organisation of both plays. To investigate the patterns of structural variation, I shall select and quantify the total frequency of interaction variables for the analysis. Such an analysis is extremely useful as it can provide the basis for a reliable structural comparison of these texts. The quantification of interaction variables will be carried out by examining the two text files directly. After, the extracted data will be computerised, tabulated (intra-play), cross-tabulated (inter-plays) and presented in tables, graphs and schemes. The readings of Hamlet and Gamlet suggest that the distribution patterns of the interactions of each main character with all characters, both main and secondary, and vice versa, as well as the relationships that are established among them are not necessarily parallel per act: intra-play and inter-plays. Moreover, it seems that the interactions are not only distributed differently but their impact is also completely dissimilar per act and per full text: intra-play and inter-plays. My hypothesis is that Shakespeare and Sumarokov probably had dissimilar views about the complexity of the relationships -revealed through the interaction patterns- among all characters, both main and secondary, and that these perspectives have led Sumarokov to somehow alter the structure of Shakespeare's original play Hamlet. In general, the key findings will show considerable distinctions between the structures of the plays per acts associated with their organisation of the social network of the characters that have connections with each other.

Khudyakova, Mariya

Panel: 3. Estudios gramaticales basados en córpora

POSSESSOR NPS AND REFERENTIAL CHOICE IN ENGLISH BUSINESS PROSE (A CORPUS RESEARCH)

The choice of an appropriate referential expression depends on multiple factors. This paper is focuses on the influence of the possessor position of a referential expression and its antecedent on referential choice. The study is based on a subcorpus of the specially designed RefRhet corpus.

Kieran O'Halloran

Panel: 2. Discurso, análisis literario y corpus

ELECTRONIC DECONSTRUCTION OF AN ARGUMENT THROUGH ITS 'SUPPLEMENT': DERRIDA AND CORPUS LINGUISTIC METHOD

A by-product of new social media is an abundant textual record of engagements - billions of words across the world-wide-web in, for example, discussion forums, blogs and wiki discussion tabs. Many such engagements consist of commentary on a particular text and can thus be regarded as electronic supplements to these texts. The purpose of this presentation is to flag the utility value of this electronic supplementarity for corpus-based, critical reading by highlighting the following: how an electronic supplement can reveal particular meanings that the text being responded to can reasonably be said to marginalize and / or repress. In turn, this can show where the text's rhetorical structure can be said to be unstable, in a state of deconstruction. Given the often large size of these supplements, knowing how to mine them with corpus linguistic software is essential. I refer to this new type of corpus-based analysis as Electronic Deconstruction. Electronic Deconstruction takes its theoretical orientations from the philosopher, Jacques Derrida, and, in particular, his idea of the supplement. We normally understand a supplement as something which is an add-on and thus outside that which is being supplemented. In contrast, for Derrida (1976), any supplement has an undecideable 'inside-outside' relation, e.g., vitamin supplements are both outside the diet in providing additional vitamins and inside the diet in replacing a lack of vitamins. I report on recent, Derrida-inspired research (O'Halloran, 2010) where I examine how a discussion forum appended to an argument in an on-line newspaper is simultaneously outside and 'inside' the argument; that is, it is a Derridean supplement. By employing statistical keyword analysis of this discussion forum supplement via WMatrix software (Rayson, 2008), using the BNC Sampler written corpus as a reference corpus, I reveal that the discussion forum carries meanings which occur as traces inside the argument, permitting a judgement that the argument seeks to marginalize / repress these meanings. Once these traces are revealed, the argument's rhetorical structure is shown to deconstruct itself. Electronic Deconstruction can be seen, on the one hand, as an intervention into the text, that is, on the basis of the discussion forum supplement as outside the argument. On the other hand, it is an 'intravention', a bringing out of meanings that already exist as traces within the argument, that is, on the basis of the discussion forum supplement as 'inside' the argument. In being simultaneously intervention and 'intravention', the analytical procedure mirrors the undecideability of Derrida's notion of the supplement. Lastly, because the procedure for locating salient concepts in the forum is statistically informed, it reduces arbitrariness in making judgements of repressions and marginalisations as well as in selecting points into the argument before going on to reveal its deconstruction.

Knörr, Garikoitz and Keith Stuart

Panel: 4. Lexicología y lexicografía basadas en córpora

THE SENSE AND SYNTAX OF 'SPEAK' AND 'TALK'

This paper presents a corpus analysis of 'speak' and 'talk'. Based on data provided by two large corpora (BNC and COCA), the aim is to point out some relevant differences in the use of these two often seemingly overlapping lemmas: the way and frequency with which they combine with adverbs (eg. 'speak quietly' vs. 'talk quietly'), the use of prepositions ('speak with/to' vs. 'talk with/to'), and their degree of productivity both in the formation of compounds and collocations and as stems (eg. 'speakable', 'talkative'). The kind of information that can be gleaned from a large corpus or several large corpora is not always to be found in dictionaries or grammar books. In particular, when using a corpus, you can see how a word behaves in its immediate context and in the larger context of the text. Therefore, the paper also includes a brief overview of the definitions and usage notes offered in the most well-known reference works and how they differ from the data provided by the corpora. Finally, we will attempt to show that the choice of a particular verb tense seems to motivate the choice of the

verb. In other words, we will try to demonstrate that there is a correlation between sense and syntax (Sinclair, 1991).

Kompara, Mojca

Panel: 4. Lexicología y lexicografía basadas en córpora

IS AUTOMATIC PRODUCTION OF DICTIONARY ENTRIES IN THE FIRST SLOVENE ONLINE DICTIONARY OF ABBREVIATIONS SLOVARČEK KRAJŠAV POSSIBLE?

The possibility of automatic production of dictionary entries in the first Slovene online dictionary of abbreviations Slovarček krajšav in Termania software is discussed in this paper. The paper presents the newly build Slovene software for dictionary production (Termania) and the possibility of automatic production of abbreviations' dictionary entries. As a first step, a demonstration algorithm has been used which focuses on the automatic recognition of abbreviations and abbreviation's expansions (Taghva 1999) in Slovene and with a restricted number of characters for each abbreviation (Kompara 2010). Further development expands the number of characters for each abbreviation to ten and takes into consideration all four types of abbreviation-expansion patterns. In the next stage, the algorithm is provided online in a demonstration version. At this stage, a random selection of Slovene text is used to verify the performance of the algorithm and to improve recognition. The upgraded algorithm is then fully capable to handle large text databases and is used on a Slovene corpus of over 60 million words. In 30 minutes, the software filters the whole corpus and provides 5,000 abbreviation-expansion pairs. The acquired data is then manually cleaned; good pairs are verified and used for production of the first Slovene abbreviations' dictionary Slovarček krajšav. For entry production the Termania software is used. Dictionary entries are divided into simple and complex. Simple entries are produced entirely automatically, complex, due to complex structures, encyclopaedic data and translations, "semi"automatically. Simple entries are mainly Slovene, covering just abbreviation, language qualifier and expansion. The abbreviation and expansion are recognised automatically by the algorithm for recognition, language qualifiers are added automatically. In simple entries we are focusing on the automatic production of nominative Slovene structures of abbreviation's expansions out of non nominative structures, as seen in example (1)

(1) AB Alzheimerjevo boleznijo (non nominative structure)

→ Alzheimerjeva bolezen (nominative structure)

Such approach is used also in complex entries. The main problem in complex entries are encyclopeadic data and transaltions for now included manually, but in the future automatically. The algorithm for automatic recognition of abbreviations and abbreviation's expansions is the link between the electronic text and the "semi"automatically produced dictionary of abbreviations. Such dictionary represents the future of electronic lexicography (Kompara 2009).

Krasnikova, Anna

Panel: 8. Los córpora y la adquisición y enseñanza del lenguaje

CORPORA AND TEACHING OF EDITING

Discussing the use of corpora for teaching we rarely mention editing. Meanwhile corpora can serve as one of the major tools for editing courses. It is possible to distinguish two main goals that are set by a teacher:

1) to teach students "to work mechanically", that is to impart them some skills and let them develop these skills to automatism;

2) to teach students to work creatively with a text, to practice critical approach and to read thoughtfully.

These two goals are achieved through different types of exercises and, accordingly, different types of corpora usage. It seems to us that the following distribution is effective: a teacher creates exercises for practicing the of editing skills, and students check their estimations and assumptions, learn to formulate and prove them.

1) Work of a teacher: creation of exercises. Editing skills depend on practice. And if you want to teach students to edit, it is necessary to have them do hundreds of exercises on different types of errors. Textbooks do not help much: while their content is enough to get acquainted with different kinds of errors, it is not enough to get hold of practical application. By means of search in language corpora it is possible to collect a material for exercises on analysis and estimation of different text aspects: language and style, logical connections, and facts.

2) Independent student work: raising of language awareness. Students often feel that there is "something wrong" with a phrase, but cannot tell what exactly is wrong and cannot explain why. They have to raise their language awareness, to prove their text estimations, and that is also where use of corpora proves to be effective.

Ktari, Imen

Panel: 7. Lingüística computacional basada en corpus

POSTMODIFIERS ACTING AS COMPLEMENTS AND ADJUNCTS IN POPULAR AND ACADEMIC MEDICAL ARTICLES: A GENERATIVE CORPUS-BASED APPROACH

Carnie (2001), following Chomsky's theory, studies postmodification, a linguistic structure that comes after the head noun to modify it, following the three levels of projection of the X Bar Theory : a minimal projection (X), an intermediate projection (X' or X bar) and a maximal projection (X'', X double bar or XP). In this paper, the focus will be laid on one of the major contributions of this theory which consists in the distinction between complements and adjuncts within the noun phrase as far as postmodifiers are concerned. Sister to the head and daughter of the single bar level, the complement is "adjacent to the head" i.e. "closer to the head than an adjunct" (Carnie, 2001: 117). Hence the complement rule

X' X (WP)

The adjunct, on the other hand, is a sister to and a daughter of a single bar level. (Carnie, 2001, p 117) and "may be freely added to any number of NPs" (Kroeger, 2005: 87).

The adjunct should follow this rule:

X' X' (ZP).

Following a qualitative and a quantitative analysis (UAM Corpus Tool), this paper seeks to investigate the relationship between the syntactic and the semantic aspects along with the frequency distribution of postmodifiers acting as complements and adjuncts in both academic and popular medical articles, adhering to a comparative corpus-based approach. . The aim of this paper is to show that postmodifiers acting as complements and are thus more "lexically specified" (Kroerger, 2005: 88) are found mainly in academic medical articles since the latter display a high level of scholarliness whereas those acting as adjuncts are more recurrent in popular articles which are considered as more narrative and closer to the casual register.

Labrador-Piquer, María José and Pascuala Morote-Magán

Panel: 8. Los córpora y la adquisición y enseñanza del lenguaje

LA LENGUA Y LA CULTURA DEL VINO EN LA ENSEÑANZA DE LENGUAS EXTRANJERAS

Aunque hay mucha bibliografía en torno al lenguaje relacionado con el vino y a su elaboración desde distintos enfoques (Lehrer, A., Hommerberg, C., Chateau, C…) en la adquisición y enseñanza de la lengua partiendo de la cultura y el léxico del vino, los estudios son escasos. La innovación de este trabajo radica en la posibilidad de ser aplicado en las Facultades y Escuelas de Viticultura y Enología extranjeras, en las que además, los estudiantes necesitan dominar otras lenguas. Debido a que la lengua y la cultura van íntimamente unidas, nuestro trabajo va a versar sobre la fusión entre ellas. En una primera fase se ha recopilado léxico, expresiones, dichos, refranes, canciones, etc. en torno al vino a partir de textos literarios escritos u orales; en una segunda, en la que estamos trabajando en la actualidad, el corpus se va a centrar en la parte cultural que abarca la historia, la geografía, el arte, la música y la literatura (popular y de autor). En este trabajo de investigación presentamos una selección de muestras de este corpus. Se destaca su aplicabilidad didáctica, ya que nos sirve de herramienta en las aulas para el aprendizaje tanto de la cultura como de la lengua, dentro del marco de la enseñanza-aprendizaje de segundas lenguas.

Lacalle, Miguel

Panel: 9. Usos específicos de la Lingüística de Corpus

THE LIMITS BETWEEN AFFIXATION AND COMPOUNDING IN OLD ENGLISH: THE SUFFIX -BORA

This paper raises the question of the limits between compounding and affixation in Old English by focusing on the suffix -bora. This form is analyzed against the wider setting of the nominal derivatives to which the suffixes -a, -e, -en, -end, -ere/-re, -icge, -estre/-istre/-ystre, -o and -u have been attached. These suffixes form deverbal derivatives, as in (ge)spreca 'spokesman' ~ (ge)sprecan 'to speak, say, utter', but the case with -bora is different, thus wi:gbora 'fighter' ~ wi:g 'strife, contest, war, battle'. The suffix -bora is a verbal element, morphologically related to the verb beran 'bear'. In this sense, Quirk and Wrenn (1994) consider -bora a suffix, whereas Kastovsky (1992) does not. The conclusion is reached that -bora represents a bound form and, as such, a suffix for two reasons. Firstly, although -bora derivatives are considerably transparent, we also come across some instances of lexicalization such as candelbora 'acolyte' and wro:htbora 'the devil'. And, secondly, -bora as a free form is extremely infrequentent. According to The Dictionary of Old English, there is a single occurrence of bora 'bearer' in the corpus.

Lobejón Santos, Sergio

Panel:1. Diseño, elaboración y tipología de corpus

EL CORPUS TRACE, O CÓMO DISEÑAR UN CORPUS Y NO FRACASAR EN EL INTENTO

El Grupo TRACE (TRAducciones CEnsuradas), formado por investigadores de diversas universidades españolas, lleva años involucrado en el estudio de la historia de la traducción en la España del siglo XX y, en particular, de los efectos de la censura oficial durante el período franquista en la traducción de diversos tipos textuales. Uno de los fundamentos metodológicos sobre los que se sustenta tal investigación es la lingüística de corpus. Tal enfoque conlleva una planificación previa en cuanto a una selección homogénea y razonada de las herramientas informáticas empleadas, a efectos de facilitar la disponibilidad digital de los textos y su acceso remoto a través de Internet. En esta ponencia se desglosan los aspectos en que el diseño del Corpus TRACE, compuesto por traducciones de diferentes tipos textuales que pasaron por el filtro de la censura oficial, ha revestido una mayor complejidad. A tales efectos, se expondrán y evaluarán las decisiones que tomadas hasta el momento, haciendo hincapié en la necesidad de establecer desde el principio tanto las líneas maestras que seguirá la confección del corpus, como la elección del software que se empleará para esa tarea. En ese orden de cosas, se dedicará un apartado a comentar los diferentes estándares y plataformas de software, tanto libres como propietarios, que se han barajado para la construcción del Corpus TRACE. Como conclusión,

se mostrará la necesidad coordinar de forma efectiva las decisiones que se tomen a nivel individual, a fin de establecer una base sólida en la fase de diseño del corpus.

López Arroyo, Belén

Panel: 5. Corpus, estudios contrastivos y traducción

WRITING COMPUTERIZED ABSTRACTS: APPLICATIONS FROM A CORPUS-BASED STUDY.

Abstracts, which constitute a secondary genre based on the Research Paper (RP), have often been the object of interlingual contrastive analysis for translation and teaching language purposes among others. However, these empirically-based, cross linguistic studies should have a central role to play in offering solutions to applied problems (Rabadán, 2008: 309). This is one of the aims of the ACTRES research group. In the present paper we intend to describe the methodology and the tools devised by the ACTRES group to bridge the transition between linguistic description and procedural information. The first step of this process was to design a small special corpus of scientific abstracts, the BioAbstracts_C-ACTRES. The macro and microlinguistic characteristics of this corpus were analyzed in order to find the most prototypical rhetorical, grammatical and lexical features of this genre. Then, we identified the "anchors" (Rabadán: in press) relevant for the native speakers of Spanish. Finally, a prototype of a writing application, the Scientific_Abstract_Generator, has been designed, aiming at helping native Spanish users who are non-linguist field experts, to write scientific abstracts in English.

López Arroyo, Belén and Martín Fernández Antolín

Panel: 4. Lexicología y lexicografía basadas en córpora

CORPUS BASED APPLICATIONS: DEFINING A BILINGUAL LEXICOGRAPHICAL AND PHRASEOLOGICAL WORK ON WINE TASTING NOTES

The present paper aims at describing a bilingual (Spanish/English) terminological and phraseological dictionary on wine tasting notes. The dictionary was thought as a lexicographical corpus-based work and designed as a communicative task according to Yong and Peng (2007); hence, the main criteria when designing and making the dictionary was the final user or the group of potential users it was addressed to. In this sense, considering the great variety of users, the dictionary has several distinctive features and further applications in different fields such as ESP teaching, Translation and Interpreting, Contrastive Analysis, Marketing, International commerce, etc. Among the distinctive features, we could point out it is a bilingual dictionary that includes definitions and examples of use; however, the most distinctive feature is that the dictionary is writing oriented (Hannay 2003), in other words it aims at helping potential users write wine tasting notes in the L2. We considered that for some users understanding how a term is used in context is as important or more as understanding its meaning. In this sense, we collected and describe the phraseological information of some of the main nouns in wine tasting notes; the user will find the linguistic structure of the main nouns used in wine tasting notes in order to be used a tool for writing them. This information is given in a separate glossary as it was not possible to include it in the dictionary entries

López Vallejo, María Á. and David Prieto García-Seco

Panel: 4. Lexicología y lexicografía basadas en córpora(Póster)

LA NECESIDAD DE UN CORPUS DOCUMENTAL HETEROGÉNEO EN EL ESTUDIO DE LA TERMINOLOGÍA MILITAR DE LOS SIGLOS XVI Y XVII

Entre las distintas novedades que se dan cita en las centurias áureas, deseamos destacar la convergencia de dos hechos:

1. la proliferación de escritos de distinta temática en nuestra lengua, como fruto del ennoblecimiento del castellano, frente a la primacía que hasta entonces lideraba la lengua latina.

2. el alumbramiento de muchas de las disciplinas técnicas y científicas y el consecuente desarrollo de las terminologías que bautizarán lingüísticamente las nuevas realidades propias de las incipientes áreas de especialidad.

El historiador del léxico que pretende dedicarse el estudio de las voces técnico-científicas que se insertan en el caudal de nuestro vocabulario en el escenario renacentista tendrá que hacerse eco de estas circunstancias y partir de una premisa evidente: la importancia de acudir a las fuentes primarias que se publicaron en aquella época y la necesidad de que en dichas fuentes se corresponda con una importante variedad textual. Así, en nuestra exposición pretendemos, a propósito del análisis de algunos ejemplos, justificar la importancia que tiene partir de una información documental variada a la hora de abordar el estudio diacrónico del léxico de una disciplina técnico científica cuyos brotes iniciales tienen lugar en los albores del siglo XVI: nos estamos refiriendo a la terminología militar. Aunque no podemos obviar la importancia de algunas de las bases de datos actualmente disponibles en la red, como el Corpus diacrónico del español (CORDE) o el Corpus del español de Mark Davies, entre otros, y la obligada consulta de las fuentes secundarias —repertorios lexicográficos, generales y específicos y ciertos trabajos que dentro de obras de mayor envergadura han abordado periféricamente el tratamiento diacrónico de algunas voces propias de la milicia—, ponemos de relieve la destacada utilidad que nos brinda la elaboración de un corpus original compuesto por la selección de textos significativos para nuestro objeto de estudio. Huelga señalar el protagonismo del que han gozado tradicionalmente las fuentes documentales de carácter literario, protagonismo que restringía la reconstrucción diacrónica a los niveles más estéticos de la lengua. Como quiera que la naturaleza de nuestros términos no podría encontrar su máxima difusión en los géneros literarios, su presencia será limitada (aunque no nula) en nuestro corpus, para cuya elaboración consideramos indeclinable la mayor diversidad posible de tipos textuales vinculados con la temática que nos ocupa: ordenanzas militares, tratados de artillería, fortificación y técnica militar, crónicas, descripciones históricas, epistolarios, diarios, memorias, billetes, etc. Además, procuraremos que en nuestro corpus base coexista la heterogeneidad en cuanto a la autoría se refiere y al grado de instrucción lingüística evidenciada en los escritos. Dentro del ideal humanístico de vincular el arte de las letras con el de las armas, hallamos a autores muy versados en la técnica de escribir crónicas y tratados según los cánones literarios imperantes. Pero junto a ellos, aparecen soldados de baja instrucción que cuentan sus peripecias biográficas en determinadas batallas. Tanto unos como otros comparten esta afición por el arte de la milicia, uno de los temas más sobresalientes del español clásico y en sus páginas darán cobijo a todos los asuntos relacionados con la guerra: armas, tácticas, maniobras, formaciones de batalla, asedio y defensa, ideales de comportamiento de los oficiales y soldados, etc. En la segunda mitad del siglo XVI destaca la publicación de este tipo de textos que pretendía testimoniar los avances que estaban teniendo lugar en las distintas materias bélicas. Habida cuenta de esta coyuntura, damos cabida en nuestro corpus a documentación archivística escasamente representada en otros corpus preexistentes, documentación que abarca desde los textos autógrafos en transcripciones paleográficas fiables a los textos impresos (en sus ediciones príncipe), desde los textos de carácter misceláneo hasta los tratados técnicos más concretos y desde los escritos destinados a un público hasta los de índole privada.

Lopez, Victoria

Panel: 8. Los córpora y la adquisición y enseñanza del lenguaje

EXPLOTACIÓN DE RECURSOS ON-LINE PARA LA CREACIÓN DE ACTIVIDADES BASADAS EN CORPUS

A pesar de que el término corpus ha estado muy de moda estos últimos años, su uso en el proceso de enseñanza/aprendizaje del inglés no está aún muy extendido, al contrario de lo que pasa en la investigación, en donde los estudios basados en corpus son habituales y casi imprescindibles. Tribble

(2000:31) y Mukherjee (2004:239) señalan que no parece que muchos docentes utilicen corpus en sus clases y, aunque estas afirmaciones fueron hechas años atrás, todavía están vigentes, pues esta situación no ha cambiado en exceso. Además, cabe señalar que es especialmente más reducida cuando se trata de lenguas diferentes al inglés. Los docentes de lenguas, tanto de niveles universitarios como de niveles inferiores son reticentes a la utilización de corpus por varias razones. La primera, porque no tienen los conocimientos y las habilidades necesarias, además de que consideran que la mayor parte de las herramientas de análisis de corpus están fuera de su alcance. La segunda razón, porque no están familiarizados con el procesamiento y análisis de corpus, en donde en muchos casos se necesitan conocimientos no sólo de informática avanzada, sino también, por ejemplo, de estadística. Por último, la tercera razón y, tal vez la más importante, porque tienen que enseñar muchas horas, normalmente con grupos grandes y preparar actividades basadas en corpus parece ser una tarea ardua que implica una dedicación temporal considerable y ellos no tienen ni el tiempo ni la paciencia para emplear su tiempo en tales actividades. En esta comunicación se van a mostrar actividades basadas en corpus para la enseñanza de lenguas que son aplicables a diferentes niveles y entornos de aprendizaje, tanto presenciales como a distancia, a partir del modelo DDL (Data-Driven Learning). El objetivo de estas actividades se centra en mejorar y consolidar tanto el aprendizaje del vocabulario y de las colocaciones, como reforzar los conocimientos de gramática adquiridos. Sin embargo, las fuentes y recursos utilizados para la creación de estas actividades basadas en corpus demuestran que se pueden evitar algunas de las razones que llevan a los docentes a no aprovechar las ventajas que ofrecen los corpus en el proceso de enseñanza/aprendizaje, dado que las actividades se realizan a partir de recursos on–line (corpus y herramientas) disponibles en Internet sin ningún tipo de coste y que hacen mucho más sencilla para el docente la tarea de elaboración y explotación de los corpus y, además, permiten también que los aprendientes realicen tareas de un modo autónomo fuera de la supervisión del docente. Asimismo, se monstrará también como un ejemplo de aplicación práctica de este tipo de actividades en un curso de secundaria.

Lozano, Cristóba and Amaya Mendikoetxea

Panel: 8. Los córpora y la adquisición y enseñanza del lenguaje

CEDEL2 (CORPUS ESCRITO DEL ESPAÑOL COMO L2): A LARGE-SCALE CORPUS FOR L2 SPANISH ACQUISITION RESEARCH

While second language acquisition (SLA) research has traditionally relied on experimental data, a new area of inquiry known as 'learner corpus research' has recently come into being resulting from the confluence of two fields: corpus linguistics and Second Language Acquisition (Granger 2002, 2004). But the contribution of learner corpus research so far has been much more substantial in description than interpretation (Granger 2004), with very little reference to current SLA debates and hypotheses (Myles 2005, 2007). We analyse the reasons why many SLA researchers are still reticent about using corpora and how good corpus design and adequate tools to annotate and search corpora could help overcome some of the problems observed. We do so by describing the design principles of a learner corpus of L2 Spanish we are compiling (CEDEL2) (Lozano 2009a) and its contribution to SLA research. CEDEL2 is a written learner corpus (L1 English – L2 Spanish) containing around 750,000 words (expected target: 1 million words) of all proficiency levels, plus a comparable native Spanish subcorpus. Data are being collected online mainly from universities and schools in USA, UK and Spain. It has been designed according to 10 corpus design principles proposed by Sinclair (2005), which distinguish it from other large learner corpora. Some advantages are:

(i) CEDEL2 is a deductive learner corpus designed to potentially answer any L2 research question concerning any linguistic structure.

(ii) CEDEL2 allows for a wide range of contrasts: it can be compared against a similarly designed native Spanish subcorpus acting as a 'control group' and against three interlanguage developmental stages (beginner, intermediate and advanced). It also allows for Contrastive Interlanguage Analysis (CIA) (Granger 1996) since CEDEL2 (L1 English – L2 Spanish) is similarly designed to WriCLE (L1 Spanish – L2

English) (Rollinson & Mendikoetxea 2010), so we can address key questions in SLA research, e.g., the source of L2 knowledge: L1 transfer, language-specific vs universal influence.

(iii) CEDEL2 includes a reliable and standardised measure of learner's proficiency, as recommended by Tono (2003) - essential to study L2 development.

(iv) For each learner, CEDEL2 contains precise and detailed background information in order to conduct research into critical period effects, language use patterns, likely cross-linguistic effects, etc. A preliminary version of CEDEL2 has already been used in published studies of L2 Spanish (Alonso et al. 2010a, 2010b, Lozano 2009b, Prieto et al. 2009). The next research steps for CEDEL2 are (i) to approach the intended target of 1 million words; (ii) to launch an online taster version; (iii) to continue the tagging of the corpus with particular reference to interlanguage phenomena (though future researchers will be able to tag any linguistic phenomena they wish); (iv) to make freely available the final version of the corpus via a dedicated webpage.

Luzón, María José

Panel: 6. Corpus y variación lingüística

DISCIPLINARY DIFFERENCES IN THE USE OF SUB-TECHNICAL NOUNS: A CORPUS-BASED STUDY

Recent research on academic vocabulary has suggested that these words have specific behaviours related not only to the genre but also to the discipline (e.g. Hyland and Tse, 2007; Martínez et al., 2009). In this research I use a corpus-based methodology to analyse how a type of sub-technical vocabulary highly frequent in academic texts (which I will refer to as "research nouns" and "discourse nouns") is used in two different disciplines (Applied Linguistics and Environmental Engineering). The purpose is to determine whether there are differences in the use of these nouns in both disciplines in terms of frequency, the lexico-grammatical patterns in which they occur, and the discourse functions associated with these patterns. The results provide corpus evidence for disciplinary variation in the frequency and collocational behaviour of sub-technical nouns. They also reveal that some of these nouns contribute to multi-word units that are part of the specific phraseology of the research paper in these disciplines. These findings suggest the need to develop discipline specific academic wordlists, which should include not only the lexical items that are relevant in a discipline, but also information on their collocational behaviour and on the rhetorical functions with which they are associated.

Macdonald, Penny, Susana Murcia, Maria Boquera, Ana Botella, Laura Cardona, Rebeca García, Esther Mediero, Michael O'Donnell, Ainhoa Robles and Keith Stuart

Panel: 8. Los córpora y la adquisición y enseñanza del lenguaje

ERROR CODING IN THE TREACLE PROJECT

This paper presents the approach to error analysis within the TREACLE project, the aim of which is to profile learner proficiency to help inform teaching curriculum design. We will introduce the error annotation methodology used on a corpus of written texts by Spanish learners of English at University level. First, we will discuss the underlying principles of the error coding scheme and then provide more details about the coding scheme itself. To ensure coders are annotating the texts in the same way, two steps were followed. Firstly, we developed a comprehensive coding criteria description giving full details as to how to code particular instances. Secondly, we performed two intercoder reliability studies to help us identify areas where coders were differing so that we could address these areas. We will present the preliminary results of the error analysis and discuss their repercussions for grammar teaching at university level.

Maiz, Gema

Panel: 4. Lexicología y lexicografía basadas en córpora

THE OLD ENGLISH VERBAL SUFFIX -LÆCAN: DICTIONARY FREQUENCY VS. CORPUS PRODUCTIVITY

The aim of this paper is to compare the corpus and the dictionary productivity of the Old English weak verbs suffixed in -læcan. The main sources for this research are the lexical database of Old English Nerthus and the online Dictionary of Old English Corpus. The assesment of productivity is based on the distinction between type-frequency (dictionary-based) and token-frequency (corpus-based). The conclusion is reached that the type-frequency and token-frequency of -læcan are very low, whereas its productivity is relatively high (except in poetry) taking into account the number of hapax legomena. Additionally, -læcan verbs are much more frequent in prose and glosses than in poetry.

Manero Richard, Elvira

Panel: 4. Lexicología y lexicografía basadas en córpora(Póster)

ELABORACIÓN DE UN CORPUS DE TEXTOS PROCEDENTES DE BLOGS PARA EL ESTUDIO DE LA CREACIÓN LÉXICA EN ESPAÑOL

Este póster muestra el modo en que se ha elaborado un corpus de textos procedentes de blogs y los fines para los que este corpus ha sido confeccionado. El objeto de estudio es el léxico propio de Internet, concretamente algunos procedimientos y características de la creación léxica en este ámbito, así como la naturaleza de los nuevos formantes y unidades léxicas que, originados o no en la Red, contribuyen a formar nuevas voces en ella. El corpus se ha realizado a partir de veinte blogs en español. Se ha elegido este formato por ser, en opinión de los estudiosos del tema, el que mejor representa la irrupción de los llamados "medios sociales de la Red". Los que componen la muestra tienen como autores a dos tipos de usuarios, lo que ha permitido comparar el tipo de léxico y procedimientos de creación léxica de ambos tipos de blogueros.

a) En primer lugar, se han incluido diez blogs cuyos perfiles corresponden a autores ("blogueros") adolescentes, preuniversitarios, de entre 14 y 17 años. Esta elección se justifica en que los procedimientos léxicos y los términos elegidos por los adolescentes quizás sean los que triunfen en el español de la Red en pocos años. Se trata de autores no expertos en la Red ni informáticos, aunque sí interesados en la tecnología y familiarizados con la comunicación en Red. Algunos de los blogs analizados han sido premiados por su calidad y todos los estudiados son de los más "enlazados" y conocidos en la Red entre adolescentes.

b) En segundo lugar, se han analizado 10 blogs cuyos autores son "superusuarios" de Internet, esto es, personas con un uso activo de la Red, de nivel avanzado, experto. Estos blogs tienen gran prestigio y reconocimiento y versan sobre tecnología, negocios digitales, comunicación, periodismo y blogosfera, temas más proclives a que aparezcan términos de nueva acuñación. La primera muestra, extensa, de textos donde se incluyen las voces estudiadas se ha recogido entre los días 6 y 20 de agosto de 2009. Las unidades analizadas en ella se han agrupado considerando el modo o modos como se utilizan en el corpus: como lexemas independientes; como formantes unidos a bases léxicas; como constituyentes de compuestos o acrónicos (aportando contenido lexémico); o como unidades que sirven de base léxica para la formación de términos por afijación; o bien, lo que es más frecuente, como elementos que presentan varios de estos valores según su contexto de aparición. Algunos de los formantes y términos estudiados son e-, i-, @, ciber, web, blog, blogging/Twitter/Internet, online y offline, wiki, .com/puntocom/puntoes, post, Mac, clic o 2.0 y 1.0. Finalmente, una vez realizado este análisis, el siguiente paso que completará esta investigación consiste en recoger, con los mismos criterios de la primera muestra, una segunda muestra de textos "colgados" en 2011. Se pretende, así, observar la evolución que, en lo que toca a los formantes y voces estudiadas, haya podido experimentar el léxico de la blogosfera en español, donde los cambios suelen sucederse con gran celeridad.

Manero Richard, Elvira and David Prieto García-Seco

Panel: 2. Discurso, análisis literario y corpus (Póster)

ELABORACIÓN DE UN CORPUS DE UNIDADES FRASEOLÓGICAS A PARTIR DE TEXTOS LITERARIOS

En nuestro póster se describe la creación de un corpus de UF a partir de un texto literario, concretamente de la novela Las ratas, de Delibes (1962). Este corpus ha sido elaborado por los miembros del grupo de investigación FRASEMIA, perteneciente a la Universidad de Murcia, con el objetivo de obtener muestras de uso literario de UF para realizar estudios de diferente índole. El corpus se construyó de la siguiente manera: en primer lugar, se localizaron las UF y se procedió a clasificarlas —en cuadros de Word— por tipos. Asimismo, estas fueron distribuidas según el personaje que las utilizaba, el número de página y capítulo donde se encontraban, el registro al que pertenecían y su contexto de uso. A continuación, se volcaron los datos en Access para poder realizar búsquedas con diferentes criterios: unidad, tipo de unidad (tal como aparece en el texto o lematizada de acuerdo con el DRAE), personaje, página, capítulo, registro al que se adscriben las unidades, sociolecto del personaje que las emite y contexto de uso. Las búsquedas que permitía esta base de datos han servido para la realización de estudios traductógicos (en inglés, francés y alemán) y un estudio en español de corte lingüístico-literario. En esta última investigación, en la que nos centramos en el presente póster, queríamos determinar si las UF que aparecen en la novela se ponen al servicio de la caracterización de personajes y de la creación del estilo del narrador. A este respecto, hemos podido demostrar con las muestras del corpus que las UF empleadas por personajes y narrador constituyen uno más de los elementos lingüísticos puestos al servicio de la creación de su discurso. La caracterización se produce, primero, por medio de la información diastrática y diafásica que arrojan las UF utilizadas, lo que ha podido comprobarse con búsquedas a partir del campo 'registro', 'sociolecto' y 'contexto de uso'. En este sentido, existe una clara diferencia entre las UF del narrador, adscribibles al registro neutro-formal, y las de los personajes, de registro neutro-coloquial. En segundo lugar, las búsquedas por los campos 'tipo de unidad' y 'personaje' han permitido observar que los personajes utilizan ciertos tipos de UF ausentes de los textos del narrador, como fórmulas rutinarias o paremias, especialmente refranes. Se trata de UF con mayor grado de expresividad y propias de la interacción social. En tercer lugar, se ha mostrado, con base en los campos 'registro', 'sociolecto' y 'contexto de uso', que entre los personajes existen diferencias en la utilización de las UF, bien por la capacidad de ciertos personajes, dado su nivel sociocultural, de emplear unidades pertenecientes a varios registros, bien por la propensión de otros al empleo de UF coloquiales o malsonantes. La caracterización, por último, también se logra por medio de la asociación de UF particulares a determinados personajes.

Marcelino, Isabel, Gaël Dias, João Casteleiro and José Martinez-De-Oliveira

Panel: 9. Usos específicos de la Lingüística de Corpus

SEMI-AUTOMATIC CONSTRUCTION OF THE UNIFIED MEDICAL LEXICON FOR PORTUGUESE

The integration of standard terminology systems into a unified knowledge representation system for biomedicine has formed a key area of research in recent years. The Unified Medical Language System (UMLS) (Humphreys et al., 1998) is the most well-known medical knowledge database, which combines the Metathesaurus, the SPECIALIST lexicon (Browne, McCray e Srinivasan, 2000) and the Semantic Network. However, the UMLS is mostly dedicated to the English language. Indeed, only a few languages are included in its core, which coverage is very limited. For instance, (Zweigenbaum et al., 2003) show that only 2% of the medical French terminology is included in the UMLS. As a consequence, many different projects have been appearing such as the UMLF (Zweigenbaum et al., 2003) for French and the efforts of the German Institute of Medical Documentation and Information to produce data for the German language for the original UMLS. But, most of the methodologies used so far to build a UMLS are based on using the original or the translated version of the MeSH (Medical Subject Headings) thesaurus , which is the most important resource of the Metathesaurus. To our point of view, in order to build a dynamic medical knowledge database, the medical language needs to be sampled by analyzing large and diversified corpora, representing diverse medical areas and genres, and by compiling existing controlled

medical vocabularies in the form of terminologies, meta-thesauri or glossaries. Indeed, although the MeSH is a valuable resource, it needs constant manual updating to follow the dynamicity of the language. As a consequence, maintaining the MeSH and the UMLS is costly, time consuming and may not reflect the reality of the medical language in due time. Moreover, it is defined based on manual indexing, which may not reflect the reality of relations between concepts as evidenced (Fellbaum, 1998) for WordNet with the famous Tennis Problem. To avoid such limitations, we propose to semi-automatically build a unified medical Metathesaurus for the Portuguese language called the UMLP (Unified Medical Lexicon for Portuguese). Our idea is first to build a unified lexicon based on electronic dictionaries, online glossaries and taxonomies (Tardelli, 2007), Wikipedia and Wiktionary. Then, based on the automatically created thesauri from online resources, we aim at constructing the Portuguese Metathesaurus. In this paper, we will specifically focus on the construction of the unified lexicon and the automatic construction of thesauri, and show how corpus evidence can improve the unification process. Our work resulted in the construction of the biggest medical unified lexicon for the Portuguese language with approximately 85,000 entries together with their respective taxonomic paths from different resources.

Marin Perez, María José and Camino Rea Rizzo

Panel: 1. Diseño, compilación y tipos de córpora

DESIGN AND COMPILATION OF A LEGAL ENGLISH CORPUS BASED ON UK LAW REPORTS: THE PROCESS OF MAKING DECISIONS

The implementation of the Bologna Reform has brought about a substantial change in the status of English as a subject in Higher Education programmes barring degrees in English studies and Translation. The new European Higher Education system aims to qualify graduates for professional competences among which the mastering of a second language, particularly English, is a must. The presence of English in current universities programmes has resulted from the choice between two possible ways of integration: the adoption of English as the language of instruction in a considerable part of some compulsory subjects, or the offer of English for specific purposes courses, as a separate subject independently of content courses. The latter is the case of Legal English incorporated into the degree in Law at the Law Faculty of the University of Murcia which the authors have been and will be in charge of teaching. It was a hard task to decide on teaching materials when first facing the subject. Legal English is a particularly obscure variety of ESP, Jonathan Swift would state in Gulliver's Travels as early as 1726 that it is (…) a peculiar Cant and Jargon of their own, that no other Mortal can understand (in Mellinkoff, 1963: 5). In addition to this, the amount of available materials, especially text books, was considerably scarce, as usually happens in other branches of ESP (Rea, 2010a). Resorting to specific corpora could have been an option, as McEnery and Wilson affirm (1996: 121): such corpora can be used to provide many kinds of domain-specific material for language learning, including quantitative accounts of vocabulary and usage which address the specific needs of students in a particular domain more directly than those taken from more general language corpora. Nevertheless, to our knowledge, the amount of written legal corpora is also reduced, and access to them, except for a few cases, is not complete. As a consequence of the scarce amount of such corpora and the methodological void derived from it, we engaged into ESP corpus design and decided to create the British Law Report Corpus (BLRC): a legal English corpus that could act as a reliable source for the development of new teaching material and further language analysis. The aim of this paper is to present the process of design and compilation of the BLRC, according to Corpus Linguistics standards as stated in Wynne (2005) for general corpora and its adaptation to specific corpora (Rea, 2010b). First, the legal corpora found are introduced; next, we give a detailed account of the design process and justify the reasons that lead to the selection of this legal genre, the mode of the texts, the organization of the corpus into different categories, the distribution of texts per category, etc.; to finish with some final remarks on further corpus applications and future research.

Marqués Aguado, Teresa and Laura Esteban Segura

Panel: 1. Diseño, compilación y tipos de córpora (Póster)

TEXSEN APPLIED TO A CORPUS OF MEDICAL TEXTS IN MIDDLE ENGLISH

Historical corpora may be used as powerful tools to investigate the development of any language, whether synchronically or diachronically, and much more so if they are annotated. On many occasions and due to phenomena such as spelling variations or the existence of declensions, for instance, annotation may be indeed an asset. In spite of the existence of computer programmes that allow the user to extract various types of information from a corpus (such as Wordsmith or Wordcrunch), the peculiarities of a Middle English annotated corpus such as The Corpus of Late Middle English Scientific Prose (currently being compiled at the University of Málaga, in collaboration with the Universities of Glasgow, Oviedo, Murcia and Jaén) are far better catered for by software tools such as Texts Search Engine (TexSEn). In our poster, we will show the process followed for the compilation of our corpus, which involves two stages: first, transcription; and second, lemmatization and tagging. Once the texts are tagged, the resulting files (in Excel spreadsheets) can be used as suitable input for TexSEn. We will also present a sample of all the potential utilities that this tool offers, such as the retrieval of word- and lemma-lists, as well as of concordances, together with the possibility of making complex searches and of building glossaries according to any user's requirements (hence showing different formats).

Marszałek-Kowalewska, Katarzyna

Panel: 9. Usos específicos de la Lingüística de Corpus

CORPUS AND LANGUAGE POLICY: IRANIAN LANGUAGE POLICY TOWARDS ENGLISH LOANWORDS

This paper will exploit the potential of corpus linguistics in investigating language policy. It focuses on assessing Iranian language policy (which is characterized by heavy linguistic purism) towards English lexical borrowings in Farsi. Two years ago the author of the paper studied English loanwords in Farsi and carried out a comparative research of technical English loanwords and their Farsi counterparts coined and approved by the Academy of Persian Language and Literature. The tool used in that study was Persian Linguistic Database – corpus of the Persian language. The results showed that in majority cases loanwords held an advantage over their Farsi counterparts. However, the majority corpus evidence was from 2002 – 2005 whereas the first Collection of Terms Approved prepared by the Academy was published in 2003. Thus, it was decided to compare the results from PLDB with the results from compiled corpus of Farsi. This paper presents a comparative corpus-driven study of certain English borrowings and their Farsi counterparts proposed by Iranian linguistic purists. These lexical borrowings belong to one semantic group – technology. The study attempts to verify the differences in usage between certain English loanwords and their Farsi counterparts. This usage relates to collocations, register and frequency. By means of compiled corpus the question about the successfulness of the Iranian language policy towards this particular semantic group will be addressed. To this end, the information about the corpus data will be presented. The aim of the study is to compare the results from the Persian Linguistic Database and corpus compiled by the author of the paper. In order to assess Iranian language policy by the means of corpus-driven study the following questions are going to be answered:

1. What are the English borrowings in Farsi? How can they be classified?

2. What is the Iranian language policy towards English borrowings?

3. What kind of data does the corpus contain?

4. What are the problems that can make the results vague?

5. Is the Iranian language policy towards English borrowings successful?

Martínez Martínez, José Manue and Iris Serrat Roozen

Panel: 1. Diseño, compilación y tipos de córpora

RECOPILACIÓN Y TRATAMIENTO SEMIAUTOMATIZADO DE CORPUS PARA EL ESTUDIO DE LA TRADUCCIÓN: PORQUE EL TAMAÑO (Y LA CALIDAD) SÍ QUE IMPORTA

El grupo ECPC ha diseñado y creado un corpus de discursos parlamentarios europeos con el fin de estudiar dicho género y la hipotética influencia de la traducción en la construcción de la noción de identidad europea. La investigación se ha restringido al Parlamento Europeo (mediante la construcción de un corpus paralelo -EN y ES- con las versiones en inglés y español) y a dos parlamentos nacionales, la House of Commons británica (HC) y el Congreso de los Diputados español CD), que constituyen sendos corpus comparables. El archivo contiene los discursos recogidos en las actas de las sesiones plenarias celebradas a lo largo de 2005 en cada una de las cámaras y alcanza un tamaño aproximado de 42 millones de tokens. ECPC es heredero directo de los estudios traductológicos de corpus iniciados por Mona Baker con el Translational English Corpus (TEC) y Stig Johansson con el English Norwegian Parallel Corpus (ENPC). Aunque el material de partida es similar al empleado para la creación de otros corpus como Opus Europarl se diferencia sustancialmente en su finalidad. Mientras que en este último caso se trata de un recurso con fines instrumentales para el campo de la traducción automática, ECPC (al igual que TEC y ENPC) tiene como principal fin investigador la descripción de la traducción. Esta diferencia se hace patente en cuanto a los criterios considerados para el diseño del corpus, su recopilación y tratamiento posterior. En nuestro caso, todo esto ha conducido a la obtención de un corpus compuesto de documentos en formato XML, que permite la incorporación de datos textuales y metatextuales. Estos metadatos posibilitan el estudio de diferentes comunidades discursivas dentro del ámbito parlamentario atendiendo a parámetros como el género, afiliación política, edad y circunscripción electoral entre otros, así como el análisis entre muestras originales y traducidas. Para llegar a este resultado, el principal desafío consistió en la creación de un corpus suficientemente representativo (Biber et al 1998, McEnery, Bowker y Pearson 2002) con unos recursos humanos y económicos limitados. Tradicionalmente en traductología la recopilación y el tratamiento de los corpus electrónicos se ha realizado de forma manual, lo que ha condenado a la disciplina a contar con corpus relativamente pequeños. Siguiendo la estela de Danielsson 2004, Hammond 2003 y Tanguy 2007, nuestra propuesta ha consistido en semiautomatizar estos procesos de modo que se mejore la eficiencia en esta fase de la investigación al obtener un corpus de gran tamaño (200 millones de tokens aproximadamente) y máxima calidad. El objetivo es que el investigador pase el mayor tiempo posible analizando su corpus y no creándolo. La particularidad de nuestro corpus radica en el género abarcado, el tamaño y la información metatextual que contiene. Estas características pueden impulsar el desarrollo del análisis crítico del discurso basado en corpus y posibilitar el estudio de la influencia de la traducción en la construcción del discurso político europeo.

Martínez Martínez, José Manuel

Panel: 5. Corpus, estudios contrastivos y traducción

¡HOUSTON, TENEMOS UN PROBLEMA... DE TRADUCCIÓN! ECPC Y TPC COMO HERRAMIENTAS DIDÁCTICAS PARA LA ENSEÑANZA/APRENDIZAJE DE LA TRADUCCIÓN

Tanto en la didáctica de la traducción (González Davies, M y Scott-Tennent 2005) como en los estudios sobre el proceso traductor (Lörscher 1991, Göpferich y Jääskeläinen 2009) o la competencia traductora (Presas 1997), se utiliza profusamente el concepto de problema (sobre todo vinculado a la producción de estudiantes). No obstante, y a diferencia de lo que ocurre con la noción de error (Castagnoli et al. 2006), el problema no ha sido estudiado de forma sistemática utilizando la metodología de la lingüística de corpus. Contar con un corpus de problemas en traducciones de estudiantes puede ser útil para (a)

identificar posibles correlaciones entre la capacidad de detección de escollos y el resto de subcompetencias del traductor (Presas 1997); (b) fundamentar empíricamente la diferencia que Nord (1991) establece entre dificultad (problema individual que puede variar entre los individuos) y problema (dificultades comunes a todos los individuos) y; (c) observar la variación entre las múltiples traducciones de un mismo texto que desde un punto de vista descriptivo puede ser útil para caracterizar rasgos "universales" en el comportamiento traductor (Castagnoli 2009). El Translation Problem Corpus (TPC) se propone cubrir este vacío. El TPC es un corpus paralelo inglés-español compuesto de intervenciones emitidas originalmente en inglés procedentes de algunos de los debates del Parlamento Europeo que conforman el corpus ECPC (European Comparable and Parallel Corpus) y múltiples traducciones al español realizadas por estudiantes de traducción e interpretación de grado (Universitat Jaume I, 2º curso del grado de traducción e interpretación) y máster (Universitat d'Alacant, máster en traducción institucional). El corpus contiene información metatextual sobre los traductores, sobre las condiciones del encargo de traducción y sobre los problemas de traducción señalados por el traductor. Para su construcción se han tenido en cuenta criterios de diseño como los apuntados por (Bowker y Pearson 2002, Granger 1998, Castagnoli et al. 2006). En primer lugar, como Granger (1998), se creó un perfil traductor de cada estudiante para lo cual se recogieron datos acerca del alumnado como el sexo, la nacionalidad, el conocimiento de la lengua origen, de la lengua meta y otras lenguas, la experiencia traductora. En segundo lugar y en cuanto a los textos, se recopiló una copia del original con los problemas identificados por cada traductor sirviéndonos tanto de un instrumento similar al IPDR (Gile 2004) como de las propuestas de González Davies y Scott-Tennent (2005). En tercer lugar, se recogieron las respectivas traducciones de los alumnos. Finalmente, tanto los textos originales como los textos traducidos se transformaron en XML y se alinearon. En el ámbito de la didáctica este material electrónico podría complementar la propuesta de Bowker (2001) para ayudar al docente en su labor evaluadora. Asimismo, podría guiar la selección y creación de materiales ricos en problemas traductores adecuados al perfil del estudiantado. También los propios estudiantes podrían beneficiarse de la consulta directa de un corpus de estas características (Florén Serrano y Lorés Sanz 2008) así como de su comparación con las traducciones profesionales contenidas en el corpus ECPC.

Mat Awal, Norsimah, Imran Ho-Abdullah and Intan Zainudin

Panel: 5. Corpus, estudios contrastivos y traducción

A CORPUS-BASED STUDY ON THE LEXICO-GRAMMARTICAL DIVERGENCE IN MALAY TRANSLATED TEXT: AN ANALYSIS OF THE RELATIVE CLAUSE MARKER YANG

Laviosa (1998) suggests that corpus-based approach is the 'new paradigm in translation studies'. Since then, various translation studies utilizing corpus-based approach have been conducted. This study uses a comparable corpus to investigate the lexico-grammatical differences of the Malay relative clause marker yang as it is one of the salient lexical items found in the corpus. The comparable corpus is made up of texts translated into Malay and texts originally written in Malay. Comparable corpus presents an opportunity to discover features that occur more frequently in translated texts or 'translation universals'. Findings on these translation universals would be a valuable tool in the teaching and training of translators.

Mateo Mendaza, Raquel

Panel: 4. Lexicología y lexicografía basadas en córpora

THE OLD ENGLISH ADJECTIVAL AFFIXES FUL- AND –FUL: A TEXT-BASED ACCOUNT ON PRODUCTIVITY

The aim of this paper is to measure the indexes of productivity of the Old English affix ful both as a prefix and a suffix. This analysis is based on Baayen's (1992, 1993) framework, which comprises different measures on productivity. The major source consulted for this analysis is The Dictionary of Old English

Corpus, compiled at the University of Toronto, although some lexicographical sources are also checked in order to obtain more accurate results. This study of productivity allows for a diachronic perspective on the evolution of these affixes from the Old English period to the present. The main conclusion drawn from this analysis is that the suffix –ful is more productive than its prefixal counterpart, which implies that more productive patterns are still maintained in Present-day English in contradistinction to the disappearance of less productive ones. These conclusions are compatible with Kastovsky's (1992) statement regarding the tendency of the Old English lexicon towards lexicalization when a given morphological pattern loses its productivity.

Melguizo Moreno, Elisabeth

Panel: 6. Corpus y variación lingüística

UNA INVESTIGACIÓN SOCIOLINGÜÍSTICA DE CORPUS EN GRANADA

En este trabajo pretendemos mostrar los resultados de una investigación fonológica de carácter sociolingüístico basada en el análisis de corpus orales recogidos en la provincia de Granada. Una investigación que se plasma en la Tesis Doctoral "Convergencia y divergencia dialectal: a propósito del habla de Pinos Puente y sus contactos con Granada" (Melguizo 2007), con la que se pretende profundizar en la formación de variedades lingüísticas que se derivan de los contactos producidos en los núcleos urbanos, como consecuencia de los movimientos poblacionales procedentes de áreas rurales. En este caso, analizamos concretamente los fenómenos de seseo, ceceo y distinción fonológica en dos muestras de población: una, formada por hablantes nacidos y residentes en la localidad granadina de Pinos Puente; y otra constituida por informantes procedentes de dicho municipio pero instalados definitivamente en Granada capital. El objeto de este estudio consiste en la comparación de ambas calas poblacionales con el fin de establecer el grado de acomodación lingüística que manifiestan los inmigrantes rurales residentes en la capital granadina. El total de informantes estudiados asciende a ciento cuarenta y cuatro (setenta y dos hombres y setenta y dos mujeres) para las dos muestras diseñadas. Cada una de ellas tiene un total de setenta y dos hablantes (treinta y seis hombres y treinta y seis mujeres). Se trata de individuos pertenecientes a tres generaciones de edad: 1ª Generación (15-24 años); 2ª Generación (25-54 años) y 3ª Generación (> 54 años); y tres niveles educacionales diferentes (sin estudios: 0-6 años; estudios medios: 7-11 años y estudios superiores: más de 11 años). La edad y el nivel educativo constituyeron las dos variables fundamentales para la estratificación de la muestra de habla. En definitiva, el objetivo fundamental de este trabajo es profundizar en el comportamiento lingüístico de los hablantes pineros tras instalarse en Granada y penetrar en la complejidad del desarrollo de los procesos de convergencia y divergencia dialectal en los contextos descritos.

Mendikoetxea, Amaya, Cristóbal Lozano and Esther Ferrandis

Panel: 8. Los córpora y la adquisición y enseñanza del lenguaje

WHY WE NEED TO COMBINE CORPUS AND EXPERIMENTAL DATA IN L2 ACQUISITION

This paper presents corpus and experimental evidence regarding the acquisition of subjects by L1 Spanish-L2 English learners. As is well known, Spanish and English differ in their setting for the Null Subject Parameter, which has been widely studied in SLA research (e.g. White (1985), Liceras (1989), Ruiz de Zarobe (1998), Phinney (1987), Al-Kasey & Pérez-Leroux (1998), Liceras & Díaz (1999), Lozano (2002), and Montrul & Rodríguez-Louro, (2006), among many others). It has been recently observed that learners do not treat all subjects alike. In particular, while L1 Spanish-L2 English learners have no difficulties in acquiring referential subjects, non-referential subjects (expletives it and there) remain problematic even at advanced stages. According to Ruiz de Zarobe (1998), once a Spanish learner of English has acquired the use of expletives, s/he is able to reset the initial parameter and adopt the target language parameter setting. Most L2 studies on the acquisition of the Null Subject Parameter are experimental. It is only very recently that researchers have started using corpora to test SLA hypothesis.

The findings reported in Oshita (2004) and Lozano & Mendikoetxea (2010) regarding issues related to the acquisition of different aspects of the Null Subject Parameter show that large and well constructed corpora and databases are powerful tools that are crucial for understanding the processes that constrain L2 production. In this study we used two L1 Spanish-L2 English learner corpora (WriCLE and WriCLEInf corpora), compiled at the University Autónoma de Madrid (see Rollinson & Mendikoetxea 2010). A random selection of texts (different proficiency levels) were annotated according to the properties of referential and non-referential subjects. A preliminary analysis of the facts confirms the hypothesis that learners show difficulties in acquiring non-referential subjects even at advanced stages. In particular, even advanced learners omit subjects in certain contexts (they use 0-subjects) and overuse it as the generic expletive, while the use of there with verbs other than be is highly limited (see also Lozano & Mendikoetxea 2010). These results are then compared with those obtained through an acceptability judgement task, in which subjects were asked to rate the acceptability of clauses containing the following subjects: it, there, 0, and a Prepositional Phrase. The results of the experimental tasks mostly match those obtained in the corpus study, so that we can talk about converging evidence, but they also show some interesting deviations, probably due to task differences.

Moerth,Karlheinz, Niku Dorostkar and Alexander Preisinger

Panel: 1. Diseño, compilación y tipos de córpora

GLEANING MICRO-CORPORA FROM THE INTERNET: INTEGRATING HETEROGENEOUS DATA INTO EXISTING CORPUS INFRASTRUCTURES

Over the past decade, the issue of Web as corpus has been discussed and studied extensively. Meanwhile, the existence of a number of very successful projects and the ever growing number of new corpora created from sources on the internet bears advocates of this new brand of NLP resources out. The number of tools that serve the purpose has steadily grown, some of these also provide web-based interfaces. The meanwhile well-established methodology of creating corpora from the Web has produced tools that allow the wholesale creation of large corpora. The software usually proceeds from so-called seeds, then crawls the Web collecting links and downloading relevant data for future reference. The most obvious area of application that comes to mind is lexicography, most software developments that have been presented are geared towards the needs of researchers looking for words, less to the reading and interpreting kind of scholars. While creating ever larger corpora has become a comparatively easy task for computational linguists, other groups of researchers who might also be interested in archiving and exploiting such data still come up against a number of difficulties that often impede smooth access to data. Our paper describes a newly developed piece of software and touches on use cases from projects where researchers need more than mere KWIC lines. It will focus on issues of interface design and key functionalities implemented in the new tool which comprise among others the selective incorporation of particular documents from the internet into a corpus and their preservation (including styles and images) allowing subsequent reading and interpretation of the text. Among the design objectives of the development project was to also enable non-technical users to archive data from the internet, to organise this data into reusable micro-corpora, to enhance data with more fine-grained metadata and to integrate them into an existing corpus infrastructure. The usability of the new tool has been put to trial in several small projects, the most important of which is a project bringing together scholars and high school students working collaboratively on racist language in online discussion forums applying methods of critical discourse analysis. The software discussed in the paper has been developed as part of a more general corpus toolbox comprising editing (corpusEditor) and access (corpusBrowser) tools. Development activities have been carried out with a strong emphasis on standards (XML, Unicode, LAF, ISOCat) and de facto standards (TEI, XCES). All the components being discussed in the paper will be freely available and published as open-source.

Mojca Kompara, Ana Begus and Elena Sverko

Panel: 4. Lexicología y lexicografía basadas en córpora

COMBINED APPROACH TO MODERN LEXICOGRAPHIC TOOLS: THE CASE OF THE FIRST SLOVENE DICTIONARY OF TOURISM TERMINOLOGY

This paper presents the first Slovene Dictionary of Tourism Terminology. In Slovene there is still no contemporary explanatory dictionary of tourism available. The only reliable explanatory sources remain foreign dictionaries of tourism. However, these dictionaries do not cover specific Slovene tourism-related terminology. That is why the production of a contemporary dictionary of tourism is essential. The paper presents the newly built Slovene Dictionary of Tourism Terminology, compiled on the basis of the Multilingual Corpus of Tourist Texts (Mikolič et al. 2008). The Corpus was compiled with the aim to draw up a Slovene-Italian-English corpus of tourist texts; to conduct analysis of these texts based on theoretical starting points of intercultural pragmatics, translation theory, critical discourse analysis and terminology, and thus to set up a platform for the compilation of a terminological dictionary of tourism. The Corpus includes 27 million words, mostly in Slovene, but also in English and Italian, thus representing a bigger multilingual LSP corpus for Slovene language (Mikolič et al. 2008). As research shows (Gorjanc 2002: 75), terminological electronic corpora represents an indispensable basis for compiling LSP dictionaries. The Dictionary of Tourism Terminology is being compiled using a newly designed software Termania (Amebis, 2010), which provides a flexible and user-friendly interface for editing dictionary entries. The dictionary currently consists of approximately 2,000 terms. In the compilation of the dictionary, the automatic and the manual approach were combined. The automatic approach was used to process corpus data and enter the processed data into Termania editing software. The most frequent tourist terms (monograms, bigrams and trigrams) were automatically extracted from the Multilingual Corpus of Tourist texts and placed in Termania software as dictionary entries. Also inserted automatically for each entry were language qualifier, grammatical and field qualifiers, examples of use and translation into English. Manual approach was then used in consecutive editing phases for correcting, complementing or adding new data for individual entries. As an example - for field qualifiers, automatic approach was combined with manual, since new fields could be added manually to the existing ones. In a similar manner, good examples and translations were checked for suitability and edited if necessary. Entirely manual approach was used for writing definitions, where editors drew upon different sources, both printed and electronic, in order to compile the definition, stating all the sources at the end of the entry. The results show that automatic approach in compiling LSP dictionaries is useful and helpful for the lexicographer but cannot replace him. A combined approach, building on the advantages of automatic and the manual approach, therefore seems the most appropriate. As shown in Humar (2004: 20-21), a good terminological dictionary is usually the result of group work which draws together the knowledge and experience of specialists from different fields. Nevertheless, the Dictionary of Tourism Terminology represents a good example of a corpus-based LSP dictionary in the electronic format, which represents an important trend of future development in the field of electronic lexicography.

Mola, Montserrat and Jordi Cicres

Panel: 8. Los córpora y la adquisición y enseñanza del lenguaje

PROGRAMACIÓN DIDÁCTICA MEDIANTE EL USO DE CÓRPORA

El propósito de esta comunicación es analizar algunas de las programaciones de enseñanza del catalán como L2 para adultos desde la óptica de la lingüística de corpus. Parece lógico que si el objetivo es que los alumnos sean competentes en el uso de la nueva lengua en un entorno real, entre los criterios utilizados para la programación de los contenidos lingüísticos de una L2 debería encontrarse la frecuencia de utilización de los distintos elementos morfológicos, sintácticos o léxicos que se pretenden enseñar. Sin embargo, la realidad de las programaciones didácticas analizada desde la óptica de la lingüística de corpus muestra que, a menudo, la secuenciación de los contenidos lingüísticos no tiene relación con su frecuencia de uso real en la lengua. Así, proponemos utilizar los córpora lingüísticos como una herramienta útil para asistir a los programadores, y que, de este modo, puedan organizar los materiales en función de criterios más realistas y más acordes con un enfoque comunicativo. En este estudio se ha utilizado el Corpus textual informatitzat de la llengua catalana (CTILC) del Institut d'Estudis

Catalans (en su parte no literaria, que consta de más de 29 millones de palabras). Por otra parte, se han analizado tanto programaciones on-line como libros de texto (Parla.cat, Fontdelcat, Itineraris d'aprenentatge del català, Programacions de llengua catalana per a l'ensenyament d'adults, Digui, digui, Veus y Passos). Uno de los casos que ejemplificamos es el de los pronombres de relativo del catalán; según el CTILC, el pronombre relativo que es el que aparece con mayor frecuencia (488.012 veces). Queda patente de este modo que dicho pronombre es un elemento muy común en la lengua, por lo que sería lógico que se introdujera en los niveles más básicos. En cambio, el resto de relativos tienen una frecuencia mucho más baja: qual (61.625), quan (45.091), què (37.943), on (31.498) qui (29.933) y quant (5.325). El análisis de las diferentes programaciones didácticas muestran, sin embargo, que en ocasiones pronombres de relativo infrecuentes se introducen en los manuales de aprendizaje antes que otros de mucha mayor frecuencia, es decir, vemos que, no solamente el criterio de introducción de los pronombres de relativo es diferente en los diversos materiales analizados, sino que este criterio, además, no guarda relación con su uso real por parte de los hablantes.

Monaco, Leida Maria

Panel: 2. Discurso, análisis literario y corpus

MODALIZING MODERN ENGLISH SCIENTIFIC DISCOURSE: A CORPUS-BASED APPROACH TO MODAL AUXILIARIES IN 18TH -CENTURY LIFE SCIENCES TEXTS (CORUÑA CORPUS)

Scientific discourse, though often considered strictly objective and hence impersonal (Hyland 1995: 33), has nevertheless demonstrated to present a significant number of epistemic modality markers, through which the authors presumably convey their (un)willingness to commit themselves to the truth of their propositions (Hyland 1998: 3). Semantic-pragmatic studies of diverse types dealing with scientific literature, both contemporary (Salager-Meyer 1994; Vihla 1999) and historical (Banks 1991, 2008; Salager-Meyer 2001; Taavitsainen 2001; Taavitsainen & Pähta 2004), appear to show that scientists normally tend to modalize their discourse when presenting their research achievements before the epistemic community, in a way that their statements might not be perceived as categorical assertions. One of such modalizing strategies is the use of modal auxiliaries conveying epistemic meanings, such as doubt, possibility, necessity, or inference (Gotti et al. 2002), all of which appear to be a recurrent case in scientific writing (Hyland 1998; Vihla 1999). The present study focuses on modal auxiliaries presenting more or less evident epistemic meanings in a corpus of twenty scientific texts belonging to the subfield of the Life Sciences (which in turn contains diverse disciplines, such as Biology, Zoology, Botany, etc.), written in English throughout the 18th century and distributed all along the said period at a rate of two samples per decade. The given texts belong to the Corpus of English Life Sciences Texts (CELiST), a part of the Coruña Corpus of Scientific Writing, the latter being an electronic collection of late Modern English scientific literature of diverse genres and disciplines, written between 1700 and 1900. The samples analyzed in the selected sub-corpus might be regarded relevant for spotting the semantic and/or pragmatic scope of the given modal auxiliaries during a period in which English was already evolving as a language of science, but, apparently, there was not yet a standard pattern for a 'scientific English'.

Moreno Ortiz, Antonio, Chantal Perez Hernandez and Rodrigo Hidalgo Garcia

Panel: 9. Usos específicos de la Lingüística de Corpus

UTILIZACIÓN DE CORPORA TEXTUALES PARA LA EXTRACCIÓN DE MODIFICADORES CONTEXTUALES DE VALENCIA PARA TAREAS DE ANÁLISIS DE SENTIMIENTO

El creciente ámbito del Análisis de Sentimiento, en su sentido más amplio, requiere de la codificación del lenguaje evaluativo en términos de polaridad/afectividad positiva o negativa, es decir, la clasificación del componente axiológico del lenguaje. Es por tanto necesario valorar el léxico de una lengua determinada y almacenar esta información codificada en bases de datos de gran cobertura para que

nuestra herramienta de análisis de sentimiento para el español, denominada Sentitext, pueda identificarla. En principio, la valoración y clasificación de palabras individuales de una lengua puede no entrañar dificultades aparentes, sin embargo, no debemos olvidar que el análisis de sentimiento implica adentrarnos en la subjetividad del lenguaje más allá de los significados individuales de las palabras, puesto que existen una serie de variables que pueden determinar la interpretación positiva o negativa de lo expresado. En este sentido, el contexto lingüístico, o co-texto juega un papel determinante, ya que términos como "enloquecer", que literalmente apunta hacia un estado mental de carácter negativo, se puede interpretar positivamente si aparece en un titular de artículo periodístico con el co-texto "Shakira enloqueció a sus seguidores." Si bien existen elementos del contexto como la ironía que son muy difíciles de clasificar sistemáticamente, existen multitud de modificadores contextuales de la valencia dentro del contexto lingüístico más inmediato a la palabra que determinan cambios en la polaridad. El ejemplo más directo puede ser la negación. Es evidente que la felicidad es un concepto positivo, no obstante la frase "no estoy contento" automáticamente invierte la polaridad del adjetivo "contento". El problema radica en que no todas las formas de inversión de la valencia son tan fácilmente identificables como una simple negación, y esto hace que sea necesario detectarlas y catalogarlas, ya sea como modificadores contextuales de la valencia (por ej., "carecer de dignidad", "vulnerar las leyes", "hacer frente a la crisis", "superar el problema"), o como expresiones multipalabra y colocaciones en las que alguno de sus componentes tiene carga afectiva, tales como "ser un rayo de luz para el enfermo", "hacer un flaco favor al progreso", etc. Además de la inversión de la valencia afectiva, como en los ejemplos anteriores, el contexto también puede modificar el grado de intensidad de la misma, bien atenuándolo, como en "moderación salarial", "ligeramente errático", "neutralizar la amenaza", bien intensificándolo: "garantizar el éxito", "acentuar el conflicto", "extremadamente eficaz", etc. Estas secuencias también han de ser detectadas y catalogadas para optimizar el rendimiento de un sistema de análisis de sentimiento. Si prestamos atención a la diversidad de construcciones gramaticales y léxicas implicadas, incluso en estos pocos ejemplos, parece obvio que la tarea de identificar, clasificar y definir estos modificadores contextuales de valencia, no es trivial. En este trabajo describimos nuestra experiencia en el empleo de corpora en la consecución de este objetivo dentro del proyecto Sentitext: una herramienta de Análisis de Sentimiento para el español.

Moreno, Veronica and Beatriz Gallardo

Panel: 8. Los córpora y la adquisición y enseñanza del lenguaje

APLICACIÓN DOCENTE DEL CORPUS PERLA: ENSEÑANZA DEL DÉFICIT LINGÜÍSTICO EN LOGOPEDIA

El corpus PerLA (Percepción, Lenguaje y Afasia) recoge conversaciones orales de personas con diversas alteraciones lingüísticas para posibilitar su análisis posterior; en la actualidad consta de 4 volúmenes de afasia, 1 de S. Williams, 1 de Trastorno por Déficit de Atención y/o Hiperactividad, 1 de S. de Asperger y 1 de lesionados de Hemisferio Derecho. Las grabaciones se realizan siguiendo el método etnográfico, y con la presencia obligada de un interlocutor clave para garantizar la validez ecológica de los datos. La transcripción se realiza según las convenciones etnometodológicas. El corpus PerLA supone una herramienta eficaz en la docencia de Lingüística en el grado de Logopedia, donde la enseñanza teórica de las características de diversos déficits se complementa y/o contrasta con el análisis de las muestras orales presentes en el corpus.

Nešpore, Gunta, Lauma Pretkalniņa, Baiba Saulīte and Kristīne Levāne-Petrova

Panel: 1. Diseño, compilación y tipos de córpora

TOWARDS A LATVIAN TREEBANK

Treebanks are among the crucial resources for the development of NLP tools. For Latvian no such a resource currently exists. To address this deficiency the development of Latvian Treebank is ongoing. As a grammatical framework for the Latvian Treebank, the SemTi-Kamols model [Nešpore et al., 2010,

Bārzdiņš et al., 2007] is used. It is a hybrid grammar in relation to dependency and phrase structure grammars that covers both synthetic and analytical forms of Latvian — a highly synthetic language with relatively free word order. In essence, the SemTi-Kamols grammar is close to the Tesnière's dependency grammar [Tesnière, 1959]. The model is based on dependency links and the notion of x-words that roughly correspond to Tesnière's nuclei. X-words were introduced as inseparable syntactic units describing analytical forms and relations other than subordination. From the phrase structure perspective, x-words can be viewed as non-terminal symbols, and as such substitute all entities forming respective constituents. From the dependency perspective, x-words are treated as regular words — they can act as head or dependent nodes in dependency relations. Manual annotation of Treebank is very laborious; therefore the tool support is crucial. As the SemTi-Kamols model is based on the dependency grammar, we have chosen to adapt the annotation tool TrEd [Hajič et al., 2001] that is used developing the Prague Dependency Treebank (PDT) [Hajič et al., 2000]. We have developed Prague Markup Language (PML) profile for the SemTi-Kamols model. PML is XML based language for linguistic annotations developed together with TrEd and acts as default input/output format for TrEd. Developing the SemTi-Kamols PML profile, the initial SemTi-Kamols grammar model has been modified, dividing the types of syntactic relations further. The scope of x-word was narrowed down to pure analytical forms (e. g., perfect tenses, complex predicates) and multi-word units (e. g., multi-word numerals). The coordination was distinguished as a separate relation: it represents both coordinated parts of sentence and coordinated clauses. This brings the SemTi-Kamols model even closer to the Tesnière's approach, where coordination (jonction) is formed by two or more homogenous nodes that have the same function in relation to the sentence. In Latvian the punctuation represents the grammatical structure of the sentence, therefore we distinguished one more type of relations — punctuation mark constructs — the relation between the punctuation mark and the unit that evokes the use of the punctuation mark. Thus we arrive at four relation types: dependency, x-word, coordination, punctuation mark construct. As a result, we have obtained a working environment for creating the Latvian Treebank manually using the extended SemTi-Kamols model and exploiting TrEd. As a proof of concept, we have annotated first 100 sentences of J. Gaarder's "Sophie's World", in lines with the project of Parallel treebank of North European languages [Sophie]. Our future plans involve integrating TrEd with the SemTi-Kamols syntax analyzer [Bārzdiņš et al., 2007] to obtain environment for semi-automated annotation process.

Nijsen, Kasper

Panel: 5. Corpus, estudios contrastivos y traducción

"THIS PAPER ARGUES = DIT ARTIKEL BEWEERT?": IS-AV-CONSTRUCTIONS IN ACADEMIC PROSE TRANSLATION

Reporting on two corpus studies involving English and Dutch academic prose, this paper examines several issues from contrastive linguistics and translation studies. It focuses on constructions that previous studies have identified as 'IS-AV constructions': Inanimate Subject – Active Verb (Master, 1991; 2006; Šeškauskienė, 2008; 2009). Typical examples are 'this paper argues' and 'this theory claims'. Such constructions appear to play a crucial role in English academic writing, but little is known about their use across languages (Low 1999). It is therefore worth investigating to what extent they are a distinguishing feature of English scientific language only or may also be spilling over into the academic prose of other languages. Such contrastive knowledge is a prerequisite for an examination of the choices made by academic translators dealing with IS-AV constructions, which may reflect (or affect) cross-linguistic or cross-generic differences, also bringing into question broader theoretical questions with respect to translation universals. To investigate these issues, and taking the English-Dutch situation as a case in point, this paper address two main questions: (1) how does the use of IS-AV constructions in English academic prose compare with their use in the same genre in Dutch; and (2) what translation strategies are commonly used by English-Dutch translators dealing with IS-AV constructions in this genre? In order to frame the corpus studies, relevant literature from the field of contrastive linguistics is discussed, as well as previous studies focusing on the use, rhetorical function and conceptualization of IS-AV constructions in an academic context. Additionally, I briefly sketch the cultural position of English in the Dutch academic world, drawing on recent reports as well as Even-Zohar's (1990) polysystem theory. Finally, Toury's (1995) theory of two major translation universals or laws, interference and

normalization/standardization, is adopted to analyse the translation strategies found, including their relation to the cultural position of English in Dutch academia. To address the first question, a comparative corpus of English and Dutch academic prose was compiled; in the second part a parallel corpus of English source texts and Dutch translations was used. Corpus analyses reveal that IS-AV constructions are used in both languages, but their frequency in English is considerably higher. Their use in Dutch, it is argued, may be due to the influence of English as the lingua franca of the academic world, and similar developments may apply to academic writing in other non-English languages.

With respect to the translation question, the findings show that a number of strategies are possible. Despite the possibilities, however, most translators chose to retain the IS-AV constructions in their Dutch target texts. This suggests that in this case the process of interference takes precedence over normalization, a finding that may be related to the cultural prestige of the English source language in this domain. To conclude, I discuss the broader implications of the results and suggest several promising avenues for future research.

Novo Urraca, Carmen

Panel: 4. Lexicología y lexicografía basadas en córpora

A TYPOLOGY OF MORPHOLOGICALLY UNRELATED ADJECTIVES IN OLD ENGLISH

The aim of this presentation is to identify the basic and derived-basic adjectives in Old English*. The former represent morphologically unrelated adjectives which do not constitute bases of derivation for other words. The latter, derived-basic adjectives, are those derived adjectives that do not have derivatives of their own. Since the formation of the adjective in Old English has drawn little attention in previous research, this study reports the results an analysis of all the adjectives contained in the lexical database of Old English Nerthus (www.nerthusproject.com), which comprises around 30,000 lexical entries along with semantic and morphological information. This analysis requires a previous study in the derivational paradigms through which all words which hold morphological relationships of a derivational nature have been isolated. Out of the 5,790 adjectives included in Nerthus, 62 basic adjectives have been identified, as well as 43 derived-basic adjectives. The conclusions of this study are twofold. On the quantitative dimension, basis and basic-derived adjectives represent a negligible part of the Old English lexicon, around 1.8% of adjectives and 0.35 of all the lexicon. On the qualititive dimension, these adjectives often reflect a lack of linguistic evidence, given that nearly one half of them are morphologically complex. The situation, therefore, is one in which reconstruction is needed in order to account for the bases of derivation of these adjectives. Therefore, this analysis contributes to an overall the explanation for the Old English lexicon in two directions. Firstly, by offering a picture of an area of the derivation of the adjective to which no previous studies have been devoted. And, secondly, by reinforcing the derivational and paradigmatic nature of the Old English lexicon.

Oncins-Martínez, José Luis

Panel: 2. Discurso, análisis literario y corpus

A CORPUS-BASED VIEW OF REPORTING FORMULAE IN DICKENS' NOVELS

As has often been pointed out, one of the distinguishing features of Dickens' style is his mastery use of the techniques of characterization (see, e.g., Page 1973, Quirk 1959; 1961; 1979; Golding 1985). Much of this success –of paramount importance in character 'individualisation' (Quirk 1961: 20)– stems from his skilful use and exploitation of the wide variety of strategies for presenting the speech of the hundreds of characters that populate his fiction. Indeed, Dickens' novels show not only one of the richest catalogues of reporting verbs in English fiction but also what is perhaps the most varied grammatical realization of the main reporting verb in fiction, said. Drawing on the classification of reporting verbs proposed by Caldas-Coulthard (1994), and with the help of ConcGram 1.0 and Wordsmith Tools 4 software, this paper presents the preliminary results of a survey of the structures that characterize Dickens' use of reporting verbs. The data come from the corpus of Dickens' novels

(circa 4.5 mil. words). The survey is at this initial stage limited to verbs reporting direct speech and, for this presentation –and for time reasons–, it concentrates on said, discussing the most typical grammatical realizations of this reporting verb, namely, said + a manner adverb (-ly), said + prepositional phrase and said + an ing- participle clause. In order to assess the idiosyncrasies of Dickens' style, the results are finally compared with those found in a reference corpus of nineteenth-century fiction (7 authors; c. 12.5 mill. words).

Orozco-Jutorán, Mariana

Panel: 5. Corpus, estudios contrastivos y traducción

EL USO INTEGRADO DE CORPUS Y MEMORIAS DE TRADUCCIÓN: CÓMO SACAR EL MÁXIMO PARTIDO DE LAS NUEVAS TECNOLOGÍAS PARA LA TRADUCCIÓN JURÍDICA

Si bien la creación de corpus comparables es un recurso ya conocido y utilizado por los traductores jurídicos, su combinación con memorias de traducción y otros recursos documentales puede aportar grandes ventajas al método de trabajo del traductor profesional. En esta comunicación presentaremos ejemplos concretos del uso de un programa (MemoQ) que integra el uso de corpus con una memoria de traducción, aplicándolo a un género textual muy específico: las licencias de uso de programas de ordenador. La idea es explicar cómo sacar el máximo partido de las nuevas tecnologías disponibles combinando el uso de corpus y las memorias de traducción para traducir, basándonos en ejemplos concretos de la traducción del inglés al español de licencias de uso de programas de ordenador, que presentan dificultades especializadas que otro tipo de tecnologías no permiten resolver de forma adecuada.

Ortega Gil, Marc

Panel: 7. Lingüística computacional basada en corpus

ANÁLISIS LÉXICO DE UNIDADES LÉXICAS COMPUESTAS

Esta propuesta se quiere mostrar cómo se realiza el análisis léxico de unidades léxicas compuestas como las locuciones, los tiempos verbales compuestos y las locuciones verbales en español; en el marco de un sistema de análisis léxico basado en un diccionario electrónico formado por 634.500 formas, simples y compuestas, y un conjunto de gramáticas y herramientas construidas tomando las máquinas de estado finito como modelo matemático [1]. El análisis de estos elementos sobre realiza sobre un corpus de oraciones anotadas léxicamente, de modo que cada unidad léxica (palabra) se anota con su correspondiente categoría léxica y sus características morfológicas, como en el caso de los verbos, nombre y adjetivos. El sistema de análisis en el que se enmarca esta propuesta se realiza sobre un corpus de oraciones anotadas léxicamente, de modo que cada unidad léxica (palabra) se anota con su correspondiente categoría léxica y sus características morfológicas, como en el caso de los verbos, nombre y adjetivos, y permite reconocer tanto formas simples como locutivas. Dentro de estas últimas se analizan tanto las que se pueden reconocer a partir de un diccionario, como p. ej. 'ministro de sanidad', como las que requieren un análisis sintáctico posterior al análisis léxico inicial para poder ser reconocidas. Este es el caso de locuciones verbales como 'dar por sentado, que puede aparecer como 'da [siempre muchas cosas] por sentado', o de los tiempos verbales compuestos en español. En estos casos el reconocimiento de la unidad léxica no puede llevarse a cabo únicamente a partir de un diccionario o de procedimientos estadísticos, [2], y se requiere un análisis sintáctico que permita identificar como una unidad las formas que constituyen la unidad léxica locutiva y anotarla con su correspondiente categoría léxica y sus características morfológicas, a la vez que los elementos, como 'siempre muchas cosas' del ejemplo anterior, se sitúan en el contexto derecho y/o izquierdo de la unidad locutiva, p. ej. '[dar/por/sentado] siempre muchas cosas', [3]. El análisis de estas unidades locutivas se realiza en el marco de un sistema de análisis basado en técnicas de estado finito (finite state methods) en el que el análisis de las locuciones y los tiempos verbales compuesto se realiza a partir de

un conjunto de gramáticas locales representadas como transductores subsecuenciales que se aplican, mediante un proceso de transducción, sobre autómatas finitos deterministas que representan oraciones anotadas léxicamente a partir de un diccionario electrónico. En esta propuesta se mostrará también como el análisis de estas unidades léxicas locutivas permite desambiguar de forma eficiente los casos de ambigüedad, [4], como el que aparece con la forma 'sentado' del ejemplo anterior, que se asocia a dos categorías distintas, verbo y adjetivo. El sistema de análisis léxico permite representar y manipular de forma eficiente los casos de unidades léxicas ambiguas, es decir, aquellas unidades léxicas que están asociadas a dos o más clases de palabra o propiedades morfológicas; y parte de estos casos se eliminan, con un margen de error prácticamente inexistente, de durante el análisis de las formas locutivas.

Ortego, María Teresa

Panel: 4. Lexicología y lexicografía basadas en córpora

LA COMPILACIÓN DE DICOENVIRO EN ESPAÑOL (DICTIONNAIRE FONDAMENTAL DE L'ENVIRONNMENT)

La diferenciación por actividades socioeconómicas favorece la diversificación lingüística y el dominio del medio ambiente cada vez cobra más importancia para la sociedad globalizada en la que vivimos. Para salvar las barreras lingüísticas y expandir los conocimientos para su divulgación mundial, los científicos y expertos necesitan mediadores interlingüísticos que transfieran la información entre lenguas, para lo que necesitan herramientas fiables en las que apoyarse durante la actividad traductora, como los diccionarios especializados. Desde el OLST (Université de Montréal – Canadá), el equipo ÉCLECTIK, liderado por la profesora L'Homme (2007), se propuso crear un diccionario fundamental en línea sobre medio ambiente titulado DiCoEnviro , que sigue los principios de la lexicología combinatoria y explicativa (Mel'čuk et al. 1984-1999, 2007). En el presente trabajo me centraré en la metodología de elaboración de los artículos que forman parte de la versión española de DicoEnviro, todavía en construcción. Hasta la fecha, se han incluido entradas relacionadas con el cambio climático, cuya información ha sido extraída de un corpus en español elaborado por Sahara Iveth Carreño Cruz y propiedad del OLST, compuesto por 85 archivos, representativo del área de especialidad. Del mencionado corpus obtenemos, primeramente, los candidatos a término con la ayuda de TermoStat Web 3.0 (Drouin, 2003), un extractor automático de términos que a partir de un corpus, extrae los candidatos a términos según criterios de especificidad. Cada candidato a término recibe una puntuación basada en la frecuencia del término en el corpus analizado y su frecuencia en otro corpus pretratado denominado corpus de referencia. Una vez que disponemos de la lista de candidatos a término, verificamos si dichos candidatos cumplen cuatro parámetros para formar parte del DiCoEnviro (L'Homme, 2008: 88-89): denotan una entidad ligada al dominio, sus actantes son de naturaleza especializada, existen vínculos morfológicos acompañados de vínculos semánticos con otras unidades que ya forman parte del diccionario y también comparten vínculos paradigmáticos. En el caso de que se cumplan los cuatro parámetros, dichos términos pasan a formar parte del diccionario. A continuación, creamos una ficha a través del programa informático Oxygen y elegimos la forma de lematización según la categoría gramatical. Con la ayuda de un analizador de concordancias automático gratuito (TextSTAT) observamos el comportamiento del término en el discurso, distinguimos las diferentes acepciones si fuera pertinente y elegimos los contextos más representativos en los que se reflejen los posibles sinónimos, la estructura actancial, las realizaciones y los vínculos léxicos con el fin de completar la microestructura del diccionario. Por último, vinculamos el término con sus equivalentes en inglés y francés, si existieran.

Palmerini, Monica and Serenella Zanotti

Panel: 5. Corpus, estudios contrastivos y traducción

A CORPUS-BASED STUDY ON THE USE OF NARRATIVE IN ENGLISH AND SPANISH YOUTH CONVERSATIONS

Recent studies have pointed at the crucial importance of narrative in the evolution of human language (Simone 2009; Lazard 2006; Victorri 2002). Narrating, i.e. telling past stories or imagining still to come or

never existed ones, is a primordial and irrepressible need in human experience, which has presumably shaped grammar at a very deep level and which appears to be an exclusive and ubiquitous property of verbal languages. As a consequence of this primeval relation, languages display a wide array of tools aimed at implementing the narrative function. The study of narrative applies to many social science fields, ranging from literary theory, history, linguistics, anthropology, psychology, sociology, art, drama, film, theology, philosophy, education and even evolutionary biological science. Linguists' attention on narrative has often focused mainly on the analysis of the complex products of long-standing literary or oral tradition. In particular, research on oral narrative has been carried out mainly on bodies of elicited personal/autobiographical narratives (cf. Labov & Waletzky 1967; Labov 1982, 1997; Gee 1991). In this study we argue, instead, for the interest of the simplest and most fundamental context where narration surfaces, namely spontaneous informal conversation. We further characterize our object of analysis by combining two different perspectives: a sociolinguistic one, which concentrates on youth language; and a contrastive one, which compares the use of narrative in English and Spanish youthtalk. The overall approach envisaged is ultimately corpus-based, in that the analysis is carried out on and through two comparable corpora of youth language, that have been both constructed at the University of Bergen: the Corpus of London Teenage Language (COLT), and the Madrid subcorpus of the Corpus Oral de Lenguaje Adolescente (COLAm). Studies carried out over the last decade (cf. Bucholtz 2011, Stenström & Jørgensen 2009, Androutsopoulos & Georgakopoulou 2003) have demonstrated the interest of youth language as a site of innovation and paved the way for further research from a wide range of perspectives. Contrastive corpus-based studies have been carried out on the Bergen corpora, which have investigated different aspects of youth language, with special reference to discourse markers (Stenström and Jørgensen 2009). And yet a model for the investigation of the forms and functions of narrative in youthspeak is still to be developed. In this contribution we intend to make a first step in this direction, presenting a corpus-based investigation on how speakers from the same age-group in two of the most spoken and influential languages in the world use and construct narrative in conversation. After outlining the basic functional and structural properties of narrative in this language modality, we will move to illustrate the contrastive analysis conducted on specific aspects of the body of data considered. We will examine, for instance, the dynamics between narration and non-narration, "narrated world" and "commented world" (Weinrich 1964), from both a pragmatic and a grammatical point of view; the quotation strategies and the other devices used by young speakers to mark the frontier between their and the others' voices; aspects of modalization, etc.

Papp, Kornélia

Panel: 4. Lexicología y lexicografía basadas en córpora

A CORPUS-BASED STUDY OF THE PROPERTY CONCEPTS KIS/KICSI 'SMALL' IN HUNGARIAN

The near synonymy of the two Hungarian adjectives kis and kicsi is examined using corpus techniques. Cognitive linguistics has witnessed a large growth in corpus-driven approaches to language structure along with a long overdue interest in lexical semantics. Two trends have emerged in the cognitive literature on the subject. Firstly, the collostructural approach (Gries & Stefanowitsch 2003, Stefanowitsch & Gries 2005, Hilpert 2006) looks at lexical constructional associations in order to identify patterns of usage and thus the meaning of the construction and second, a multivariate technique (Gries 1999, Heylen 2005). This study considers both approaches and seeks to explain the difference in the usage of the adjectival alternation in Hungarian. The two property words in question, kis (e.g. kis ház 'small house') and kicsi (e.g. kicsi ház 'small house') are analysed within the noun phrase. The adjective kis is typically associated with attributive use, while kicsi has traditionally been identified as its predicative counterpart. There has been no corpus-based investigation into the alternation of the above mentioned adjectives in attributive position. The presentation deals with the different adjectival senses of these primarily size-related adjectives in combination with the corresponding noun senses. The study is based on the Hungarian National Corpus, where some 500 examples of each forms are annotated for semantic usage features. The semantic features consist of lexical semantic features of both the modifier and the noun. Collocational and correspondence analyses are then used to look for multivariate patterns in the usage, relative to semantic features. The results clarify the lexical constructional

interaction as well as outline a multidimensional map of the usage. This allows us to understand the lexico-grammatical meaning that produces the apparent variation.

Pennock-Speck, Barry

Panel: 6. Corpus y variación lingüística

VOICE-OVERS IN BRITISH TELEVISION ADS: A CORPUS ANALYSIS OF A WRITTEN-TO-BE-SPOKEN GENRE

The analysis of a corpus of voice-overs I will be presenting today is an integral part of a larger corpus of television ads compiled by the MATVA (Multimodal Analysis of TV Ads) group, which is made up of 636 day-time television ads aired on ITV1 on the 24th and 35th of June 2009 from 8.00 a.m. to 6 p.m. The corpus as a whole consists of a detailed description of the ads, a transcription of all the voice-overs, on-screen text, testimonials and dialogues, as well as an in-depth description of the para- (Poyatos 1993) and extra-linguistic elements of each commercial. I chose ITV1 as it is the most popular of British commercial TV channels but the two days chosen were done so randomly. Any day-long corpus of TV ads contains many repeats of the same ad –up to 30 in the case of Sky TV– and taking them all into account is important in some types of analysis. However, for corpus analysis one of each ad was deemed appropriate and so repeats were eliminated leaving 277 ads. 200 of these featured voice-overs. Although the layperson's term voice-over is well known, my definition is more restrictive as it only includes totally disembodied voices (Pennock-Speck & del Saz-Rubio 2009), thus excluding voices belonging to actors who appear at some time in the ad–these are included as testimonials and dialogues to be analyzed elsewhere. Unlike other qualitative analyses of voice-overs I have carried out in the past, here I will eschew the para- and extralinguistic characteristics of the ads and concentrate on the actual verbal messages the voice-overs are a vehicle for. One of the reasons for this is to discover, employing quantitative methods, the common lexical elements of British TV AD voice-overs. Using Wordsmith I have discovered that there are grammatical and lexical elements that predominate in my corpus. With regard to the grammatical elements, once items such as articles and prepositions have been excluded, 'you', 'your', and 'we' and 'our' and 'can' and 'just' are the commonest. Subsequent qualitative analysis has shown that the frequency of the pronouns point to the presence of positive politeness strategies of inclusiveness. The discourse analysis of the word "just" shows that its most frequent use is as a hedger, that is, a negative politeness strategy. The most frequent lexical items are 'now', 'free' and 'new'. The import of this research, apart from the findings we have made, is made more significant due to the dearth of corpora featuring TV ad voice-overs (Leech, 1996; Costa et al. 2005). The corpus analysis I will describe in this paper is only the first part of an analysis which aims to compare our TV ad voice-over corpus with both spoken and written discourse as my written-to-spoken genre partakes of both.

Perea, Maria-Pilar

Panel: 6. Corpus y variación lingüística

UN CORPUS DE DIETARIOS DE VIAJES: LOS LÍMITES ENTRE EL DIALECTO Y EL IDIOLECTO

La edición de todos los dietarios de viajes que el dialectólogo mallorquín Antoni M. Alcover (Manacor 1862 - Palma 1932) publicó entre 1900 i 1923 ha dado lugar a un corpus de carácter biográfico y documental que sobrepasa el millón de palabras. Las fuentes provienen mayoritariamente de los relatos extraídos de los 14 volúmenes de la primera época del Bolletí del Diccionari de la Llengua Catalana (1901-1926) y de las narraciones que aparecieron en publicaciones periódicas como el Diario de Mallorca (1901-1902) y La Aurora (1912 y 1913). El concepto "Dietario" congrega no sólo los ocho relatos que llevan esa denominación, sino también las crónicas de las impresiones obtenidas en las diversas excusiones que Alcover efectuó y que reciben nombres como "impresiones de viaje", "excursiones" o "escapadas". El objetivo de los viajes del dialectólogo era estudiar las formas vivas de la lengua para obtener materiales que le permitieran redactar su famoso Diccionari català-valencià-balear, pero los dietarios contienen también numerosas descripciones de los lugares que visitó, presentes

especialmente en las narraciones de los viajes realizados en el extranjero. Adicionalmente, algunos textos reúnen comentarios sobre las formas dialectales usadas en las localidades donde llevó a cabo encuestas, tanto desde el punto de vista fonético como morfológico y sintáctico. Este hecho incrementa la dificultad en el etiquetado de los materiales. Ueda y Perea (2010) presentaron un método de lematización del tomo V del Bolletí del Diccionari de la Llengua Catalana (1908), que contiene el "Dietari de la meua exida a Alemania y altres nacions durant l'any del Senyor 1907". En este estudio se presentan las dificultades que posee un corpus dialectal de estas características y se analizan los elementos fonéticos y morfológicos más destacables, propios de la variedad mallorquina, que experimentan variación a lo largo del período cronológico que abraza la publicación de los dietarios (1901-1923). Es el caso, por ejemplo, de la aparición de soluciones ieistas (rondaia~rondalla), con relación a la fonética, o el uso de formas pronominales demostrativas o posesivas alternativas (aqueis~aqueix~aquest o meua~meva), con relación a la morfología. Además de la caracterización dialectal, que el narrador adapta a los potenciales lectores de sus textos, el corpus permite también definir unos rasgos idiolectales que caracterizan la escritura del autor.

Gutiérrez, Camino, and Julia Alonso

Panel: 7. Lingüística computacional basada en corpus

THE TRACE CORPUS ALIGNER: DEVELOPING A NEW ELECTRONIC TOOL FOR LANGUAGE RESEARCHERS

This presentation aims to introduce a tool that builds a bridge between new technologies and the study of source texts and their translations. Nowadays, many aligner applications can be found in the market, but they can barely fulfill researchers' expectations, rarely satisfying all their needs. With this scenario, our goal is to develop an application that is useful and usable for researchers. By creating this software, functions such as tagger, aligner, and results screen are intended to become approachable from a single interface. The application offers several options, which are based on the needs of the TRACE project (University of León). This project is devoted to the study of the translation and censorship of different text types (narrative, theatre, audiovisual, poetry) during Franco's regime. The software already available offers alignment by paragraphs or sentences, which is not useful in the study of, for instance, theatre or audiovisual works since these texts are structured into speeches and annotations. Our goal is to develop standardized software that can be used to solve these problems, therefore making possible this type of research. Another inconvenience found in the linguistic field is the uncommon use of computer standards. This problem is quite relevant, so part of our presentation is devoted to explaining concepts such as XML, TEI or TMX, which are important standards used in our application. Thanks to these standards, intermediate and final files generated by the application can be exported, being portable and accessible for other tools we may need.

Piotr Pakuła, Łukasz

Panel: 2. Discurso, análisis literario y corpus

'CIVIL PARTNERSHIP' AND 'GAY MARRIAGE' IN CONTEXT

The question of identity has enjoyed wide interest in various fields of contemporary social sciences (e.g. du Gay et al 2000). Recently, a global shift from scrutinising linguistic differences between members of diverse social groups (e.g. Labov 1966, Trudgil 1974, Lakoff 1975, Spender 1980) to examining more abstract socio-linguistic means of expressing and describing any of the identities an individual assumes – i.e. discourses - can be noticed (e.g. Baker 2005, van Dijk 2005, Litosseliti 2006). A more recent strand of research in this field takes advantage of the blend of CDA (Critical Discourse Analysis) and corpus linguistics, as the latter "[…] can help reduce researcher bias" (Mautner 2009: 123; see also Baker 2006). However, little attention has been devoted to the discursive representation of a relationship that a member of a socially stigmatised group enters. One in-depth study done in this area is Bachmann (2011), who examined discourses surrounding the concept of 'civil partnership' as represented in the

British parliamentary debates at the time when it was undergoing legislation, i.e. 2004. Yet, because public opinion is informed mainly by the media, it was thought that investigating newspapers as one of the most profoundly opinion-shaping means might be of particular relevance. This study aims to partially fill this gap by examining different ways of talking about:

• the process of legislation of civil partnerships,

• how civil partnerships work in practice in the UK,

• and the struggle for the legal recognition of the institution of gay marriage

as represented in the most popular British newspapers published between 2000 and 2010. To this end, a corpus of c. 6 million words has been compiled; the British National Corpus served as the reference corpus for deriving keywords in the newspaper corpus. In contrast to the methodology employed in Baker (2010), no classificatory attempt has been made with respect to the traditional broadsheet/tabloid division; the categories of newspapers employing similar discourses pertaining to the subject matter emerged as the result of the analysis. The quantitative analysis was performed using WordSmith 5, then a qualitative analysis followed in order to strive for a better understanding of the keywords and their collocations. Phenomena, including nominalisation, metaphor and metonymy, were taken into account as well. Moreover, a contrastive analysis of contextualised key phrases – civil partnership and gay marriage - is presented.

Potemkin, Serge

Panel: 4. Lexicología y lexicografía basadas en córpora

SENTIMENT EXTRACTION FROM THE BILINGUAL CORPUS

In recent years, sentiment analysis has attracted considerable attention. It is the task of mining positive and negative opinions from natural language, which can be applied to many natural language processing tasks, such as document summarization and question answering. Sentiment analysis both at document and sentence level rely heavily on word level. The hypothesis is that, given the semantic orientation SO of relevant words in a text, we can determine the SO for the entire text. This paper explores methods for generating subjectivity analysis resources in a new language by leveraging on the tools and resources available in English. We focus our experiments on Russian, selected as a representative of the large number of languages that have only limited text processing resources developed to date. Note that, although we work with Russian, the methods described are applicable to any other language, as in these experiments we (purposely) do not use any language-specific knowledge of the target language. Certain semantic orientation lexicons have been manually compiled for English—the most notable being the General Inquirer (GI) [Stone et al., (1966)]. However, the GI lexicon has orientation labels for only about 3,600 entries. The Pittsburgh subjectivity lexicon (PSL) [Wilson et al., (2005)], which draws from the General Inquirer and other sources, also has semantic orientation labels, but only for about 8,000 words. The latter lexicon was used as the seed sentiment lexicon for further processing. The translation of sentiment information has been the topic of multiple publications. Some methods simply use bilingual dictionaries to translate an English sentiment lexicon. The other methods are based on parallel corpora. The source language in the corpus is annotated with sentiment information, and the information is then projected to the target language or vice versa. Problems arise due to mistranslations. Machine translation also was used for multilingual sentiment analysis. Given a corpus annotated with sentiment information in one language, machine translation is used to produce an annotated corpus in the target language, by preserving the annotations. The original annotations can be produced either manually or automatically. We use a collection of Internet blogs about new books in Russian. Each opinion in the blog is manually annotated [Zagibalov, (2010)]. This collection was translated into English using Google MT engine. Then the bilingual space techniques was applied to derive a mapping of the Russian source sentence (SS) to the English target sentence (TS). The most probable mapping defines the true matching of word pairs and multi-word fragments [Potemkin, (2010)]. The Russian words that correspond to the seed semantically oriented English words are

included in the Russian seed sentiment lexicon. Afterwards this lexicon was compared to the hand-crafted list of Russian semantically – oriented words. The advantage of this approach in comparison to the direct translation of English seed lexicon into Russian using dictionary consists in disambiguation of multiple translation equivalents.

Prieto García-Seco, David and María Á. López Vallejo

Panel: 4. Lexicología y lexicografía basadas en córpora (POSTER)

CONFECCIÓN DE UN CORPUS DE FORMACIONES LÉXICAS OCASIONALES PROCEDENTES DE LA LITERATURA DEL SIGLO DE ORO

El póster que presentamos muestra de qué modo se está llevando a cabo la elaboración un corpus de formaciones léxicas ocasionales pertenecientes a diversos autores españoles de los siglos XVI y XVII. Se trata de voces inventadas, frecuentemente de un solo empleo, tales como atalegar, bosqueril, gobernadoresco (Cervantes); chirimista, chupamadera, protonecio (Góngora); idolicida, nocturancia, pintamentiras (Lope de Vega); angelicar, desnarcisar (Cascales); archipobre, pretenmuela, protocuerno (Quevedo); cuellicida, hombrituerto, rostriamargo (Ruiz de Alarcón); armiñar, quijotista (Villegas); asacristanado, mesonil, zurraverbos (López de Úbeda); frasificar, unovolante (Calderón), etc. En primer lugar, queremos hablar de cuáles han sido las fuentes en que nos hemos basado para la compilación de las palabras que forman nuestro corpus. Mostramos entonces que nos hemos valido tanto de fuentes primarias (los propios textos literarios en que ocurren dichos términos) como secundarias. Con estas últimas nos referimos principalmente a los diccionarios que acogen en sus columnas, con o sin indicación de la procedencia, este tipo tan singular de voces, entre los que destacan el Diccionario de autoridades (1726-1739), el Diccionario castellano (a1767) del jesuita Esteban de Terreros, el Nuevo diccionario de la lengua castellana (1846) de Vicente Salvá y los dos diccionarios históricos de la Real Academia Española (1933-1936 y 1972-1996). También desearíamos exponer las características que presenta nuestro corpus, como el número de voces recopiladas hasta la fecha, y fundamentalmente las diferentes informaciones que permite recuperar la búsqueda en Access. Entre otros datos, de cada una de las palabras se ofrece el mecanismo de formación léxica, diversas anotaciones pragmáticas, el autor, la obra y la datación. La investigación que estamos realizando pretende alcanzar una serie de objetivos, de los que podemos adelantar algunos. El estudio de las palabras que componen el corpus pone de manifiesto que las formaciones léxicas ocasionales presentan una serie de rasgos comunes referidos al nivel de lengua, a su capacidad o incapacidad para ingresar en el vocabulario común, a su vinculación con el contexto literario o a los propósitos que motivan su creación. Asimismo, uno de los objetivos principales que persigue nuestro trabajo es estudiar cualitativa y cuantitativamente los procedimientos de creación léxica empleados por los escritores del Siglo de Oro, tanto los más conocidos, como la derivación y la composición, como los menos productivos y por tanto apenas estudiados, como sucede con la formación de voces por falsa segmentación (pretenmuela < preten-diente; caifascote < anas-cote).

Prommas, Pansa and Kemtong Sinwongsuwat

Panel: 8. Corpus, adquisición y enseñanza de lenguas

A COMPARATIVE STUDY OF DISCOURSE CONNECTORS USED IN ARGUMENTATIVE COMPOSITIONS OF THAI EFL LEARNERS AND ENGLISH-NATIVE SPEAKERS

This study examines the use of discourse connectors (DCs) in argumentative compositions of Thai- and English-native college students. 24 compositions were collected from third-year English major students in the Faculty of Humanities and Social Sciences at Thaksin University, Songkhla Campus whereas 20 compositions of English-native students at University of Michigan were retrieved from the Louvain Corpus of Native English Essays (LOCNESS). Following the taxonomy adopted from Halliday & Hasan (1976), Biber et al. (1999), and Cowan (2008), 140 DCs found were classified into 8 categories: addition,

enumeration and ordering, exemplification and restatement, concession and contrast, cause and result, summation, stance, and topic shift. Findings revealed that both groups of students shared similar characteristics with regard to the types of DCs employed in their essays, but with different degree of occurrence. Despite a wide range of DCs, the Thai learners, similar to the native speakers, employed a rather small cluster of DCs in their argumentative writing. And, but, because, for example and also were mostly found in the compositions of the two groups. In terms of syntactic distribution, the Thai learners had a tendency to employ the top five DCs inter-clausally as coordinators followed respectively by conjunctive adverbials and subordinators while the native speakers used them mostly as conjunctive adverbials in sentence-initial, medial and final positions, followed by coordinators and subordinators. Although both groups used these DCs in similar functions, preliminary findings suggest that the learners are more familiar with the inter-clausal rather than the intra-clausal use of DCs, associating them with clause-linking rather than intra-clausal devices, and the learners apparently had difficulties with such DCs as but, part of which can be attributed to the influence of the native language.

Quintana Toledo, Elena and Margarita Esther Sánchez Cuervo

Panel: 4. Lexicología y lexicografía basadas en córpora

AN APPROACH TO TYPES OF MODALITY IN THE INTRODUCTION AND THE CONCLUSION SECTIONS OF COMPUTING RESEARCH ARTICLES

The scientific research article comprises several parts with a different purpose. As an independent genre, it is currently assessed from several perspectives that take in lexical, grammatical and rhetorical features, among others. In this study, we seek to identify the most frequent modal auxiliary verbs encountered in the introduction and the conclusion of the scientific research article. In the introduction, the context of research and subject matter are described; it can present a summary or overview of the author's position. In the conclusion, the research contribution to the field of study is usually revealed. This shows the logical outcome as devised in the introduction. Modality can be defined as the expression of the interpersonal function of language. It concerns the way in which the author is going to project his/her attitude into his/her texts (Hyland, 2000). It also refers to how we orientate, shape and measure utterances in discourse. Furthermore, modality is related to that part of language that allows us to connect our expressions of belief, attitude and obligation with what we say and write. It includes markers of the varying degrees of certainty that we have about the propositions we transmit, and of the types of commitment or obligation that can be attached to our utterances (Simpson 2004: 123). For our study of the prevalence of several types of modality, we will regard those related to the speaker/writer's expression of volition with "will", and his/her ability to carry out the event designated with "can". We will also consider the speaker/writer's assessment of the communicated proposition with instances of epistemic modality. This encodes diverse degrees of certainty as regards its validity. For example, we encounter high certainty or necessity ("must", "cannot"), medium certainty or probability ("will", "would", "should"), and low certainty or possibility ("may", "could") (Arrese, forthcoming). Some preliminary conclusions indicate a dissimilar use of modal auxiliary verbs in the initial and final segments of the scientific research article. In the introductory section, authors manifest their impending decisions by utilising expressions of volition and intention with the modal auxiliary "will". They are also concerned with their ability to perform the intended investigation with instances of "can". In the concluding section, however, the epistemic predominance of modal verbs suggests medium and low certainty. For example, in the utterance "(…) Slices above will be unaffected, and slices below in objective will be unaffected if dominating in the other objectives", the writers predict a positive result based on their preceding research about algorithms. This paper is part of the research project "Evidentiality in a multidisciplinary corpus of research papers in English" at the University of Las Palmas de Gran Canaria. The corpus for this study includes up to twenty computing articles covering a time span that goes from 2004 to 2008. The criteria for the selection include the impact index, year of publication, and sociological aspects. The methodology is both quantitative and qualitative.

Ramírez Polo, Laura

Panel: 1. Diseño, compilación y tipos de córpora

MATVA: A DATABASE OF ENGLISH TELEVISION COMMERCIALS FOR THE STUDY OF PRAGMATIC-COGNITIVE EFFECTS OF PARALINGUISTIC AND EXTRALINGUISTIC ELEMENTS ON THE AUDIENCE OF ENGLISH TV ADS.

The structure of television commercials, the soundtrack, voice-overs, actors' accents etc. are not the result of random decisions. Rather they are chosen with a purpose in mind, that of maintaining the product in the public eye or persuading the audience to buy the product. Attesting the complexity of this type of texts, the MATVA research group (Multimedia Analysis of TV Ads) devised the creation of a database with commercials from the UK, aiming at constructing a valuable resource for the study and analysis of paralingistic and extralinguistic elements of TV ads and well as their pragmatic-cognitive effects on the audience. The following paper addresses the difficulties and decisions made in the design and construction of the database. In the first place, we address some of the theoretical questions we have faced in the conceptual design. Our objective was to create a database as a special speech corpus with an analysable textual component made up of the transcriptions of the ads. To begin with, we tackle the criteria established by the Eagles Spoking Language Working Group for the acquisition of data. Further, we discuss the criteria defined by Sinclair (1996) within the EAGLES initiative to create corpora: quantity, that is, the size of the corpus; quality or the authenticity of the corpus; simplicity or the format in which text is stored; and documentation or the metadata that must accompany the text corpus. Besides, we consider the main aspects for designing a corpus dealt with by Tourruela & LListerri (1999): its goal(s), limits and the type of corpus. Finally, we address the notion of representativity with respect to our collection of ads. We then introduce some practical issues regarding the metadata that accompanies each advertisement: the classification schema used to organize the commercials as well as the different variables that constitute the database: product types, ad duration, song lyrics etc. We also explain the markup system developed in order to annotate the transcriptions of the commercials, which was conceived with the goal of subsequent para- and extralinguistic analysis. Finally, we mention some technical factors such as the platform used to store the data as well as the structure of the data. We end with some conclusions about the extendibility of the corpus and its practical applications.

Ramon, Noelia

Panel: 5. Corpus, estudios contrastivos y traducción

'WELL' IN SPANISH TRANSLATIONS: EVIDENCE FROM THE P-ACTRES PARALLEL CORPUS

A particle such as the English form well is multifunctional. This English adverb can carry meanings related to manner, but also to degree or intensification. In addition, well is often grammaticalized into a discourse particle, especially in dialogue, and this requires a particularly careful treatment in the case of translations, as discourse particles do not carry easily definable meanings. Previous studies on the English particle well (Aijmer & Simone-Vandenbergen 2003, Johansson 2006) have shown that the translation of this item into other languages is far from straightforward, as there are many different correspondences and a high degree of omissions. The translations of the English form well have been studied in the cases of Norwegian, Swedish, Dutch, German and Italian, and this paper aims at expanding the analysis considering translations into Spanish. The study will focus on the translations of well as it appears in an English-Spanish parallel corpus, which will provide the empirical material for the analysis. The ACTRES project (Análisis Contrastivo y Traducción English-Spanish) is a long-running research endeavour currently in progress at the University of León, Spain, studying English and Spanish from a contrastive perspective and with translation-oriented applications in mind. Within this larger framework was built the P-ACTRES corpus (Parallel-ACTRES). This corpus contains about 2.5 million words of contemporary English texts and their corresponding translations into European Spanish. Various registers are represented (fiction, non-fiction, press, miscellanea) and all English texts have been published in the year 2000 or later, thus representing the current state of the language. The translations have all been published in Spain by a wide variety of different translators, thus also representing current trends in translational norms in this particular target language. The corpus-based methodology

employed will consist of a careful analysis of the cases of well in the English section of the corpus, followed by a detailed study of the various translational options identified for each function or meaning. The aim of the study is to provide an inventory of translation solutions available in Spanish for the various functions of well in English original texts, in particular with regard to the use of well as a discourse marker. The trends observed in the options taken most frequently will provide useful information in the field of translator training as well as in translation practice.

Renau, Irene and Rogelio Nazar

Panel: 4. Lexicología y lexicografía basadas en córpora

ANÁLISIS CUANTITATIVO DEL USO REAL DE LOS VERBOS PRONOMINALES ESTRICTOS DEL CASTELLANO UTILIZANDO UN CORPUS DIACRÓNICO (GOOGLE BOOKS)

Los verbos pronominales inherentes (también llamados «puros», «estrictos», «intrínsecos» o de otros modos) son aquellos que no pueden prescindir del pronombre que los acompaña: «Si de algo puedo jactarme es de haber trabajado con intensidad» (así igual en fugarse, atreverse, etc.). Este tipo de verbos es común a todas las lenguas románicas. La obligatoriedad del pronombre en estos casos ha sido considerada por algunos autores como una especie de rasgo morfológico, sin posibilidad, por tanto, de ser analizado desde el punto de vista gramatical e incluso semántico. El verbo y el pronombre se consideran estructuras lexicalizadas, «opacas» sintácticamente y con un significado propio sintético, no analítico. Este trabajo se propone estudiar los verbos pronominales inherentes del castellano con la ayuda de varios corpus. Se parte de la evidencia de que, en varios diccionarios actuales, estos verbos han sido lematizados en algunos casos en su forma pronominal (jactarse) y en otros en su forma no pronominal (jactar). Así, el criterio lexicográfico del DRAE admite la forma transitiva jactar (relativamente común en el castellano de los siglos XVII y XVIII), considerada «anticuada» por dicho diccionario, mientras que en el DEA, por ejemplo, estos usos no se admiten y, por tanto, se ofrece únicamente jactarse. Los motivos por los cuales ocurren estas divergencias se extienden, además, a diversos rasgos propiamente lingüísticos (semánticos, argumentales, de variación...). En este estudio se seguirán los siguientes pasos: en primer lugar, se recogerá la descripción de esta clase de verbos tal como es presentada en las últimas gramáticas del castellano (GDLE y NGLE) y en algunos estudios dedicados al uso de se (Sánchez López 2002, entre otros); en segundo lugar, se obtendrán los lemas pronominales de tres diccionarios generales de lengua castellana y se compararán para cotejar la proporción de casos en los que estos verbos se registran en los tres diccionarios de manera idéntica frente a los casos en que difieren; por último, se realizará una búsqueda en varios corpus sincrónicos y diacrónicos de las formas que, en alguno de los tres diccionarios, hayan sido lematizadas como intrínsecamente pronominales y, por cada diccionario, se separará el conjunto de estos verbos cuyo uso estrictamente pronominal puede documentarse tanto en un corpus diacrónico (siglos XVII-XIX) como en uno sincrónico (siglos XX-XXI). Para la investigación diacrónica se partirá en especial del corpus de libros ofrecido por Google N-grams Viewer (Michel et al., 2010), un corpus diacrónico del castellano de dimensiones superiores a lo visto hasta el momento en trabajos de lingüística. Este corpus permite buscar formas desde el siglo XVI hasta el año 2008, y ofrece la frecuencia de uso en cada época. La explotación de Google N-grams se realizará de forma automática, pues el sistema posibilita descargar el índice de enegramas del corpus completo. El objetivo del estudio es triple. Un primer objetivo es puramente lexicológico: se pretende que ayude a avanzar en el establecimiento de las características semánticas y argumentales de los verbos analizados. Un segundo objetivo es metalexicográfico, ya que ofrecería una nueva perspectiva sobre la forma en que los diccionarios representan estos verbos. Por último, se trata también de observar los cambios hacia la pronominalización de verbos a través de cuatro siglos, lo que, desde una perspectiva aplicada, puede ofrecer soluciones lexicográficas para la representación de estos verbos acorde con los datos empíricos.

Rettore-Totaro, João Henrique

Panel: 3. Estudios gramaticales basados en córpora

MENSURACIÓN DE LA VARIABILIDAD ESTRUCTURAL EN CORPORA ROMÁNICOS MEDIEVALES Y MODERNOS

El objectivo de este trabajo comprende la evaluación cualitativa de la eficacia de medios de identificación y análisis de la variabilidad estructural, específicamente en los niveles morfológico y sintáctico. La dispersión de formas, claramente dibujada a partir de la aplicación de métodos estadísticos a bases de datos textuales codificados, determina las posibles nuevas combinaciones de elementos gramaticales en estados evolutivos posteriores. La hipótesis central de la investigación está estrechamente relacionada con la idea de que la variación alcanza expresividades diferentes en la historia de los sistemas linguísticos, alternando períodos interdependientes de mayor diversidad , a veces acompañados de variación estable, y otros de alta velocidad de cambio. Dicha verificación puede ayudar a caracterizar los estados sincrónicos sucesivos en términos de grados de libertad estructurales, según su naturaleza potencial o realizada y las frecuencias relativas de sus variantes. El tratamiento gráfico comparativo de la frecuencia de ocurrencia de variantes en cortes sincrónicos es uno de los métodos disponibles; sin embargo, la correlación estadística entre las curvas de tendencia obtenidas para las bases de datos y el estudio del cambio de inclinación de estas mismas curvas ofrece parámetros numéricos que pueden ser utilizados a) para observar tipos de comportamiento diacrónico de los sistemas lingüísticos y b) para discernir similitudes y particularidades en la historia de estas lenguas (Johnson 2008). Ensayos con bases de datos de portugués y español medievales muestran una fluctuación en torno a padrones definidos por las normas gramaticales de cada época: los cortes sincrónicos revelan que cada sistema tiene sus propias características de variación interna y selección diacrónica formas. Una vez que la investigación de la tasa de implementación de las variaciones tipológicas puede demostrarse individualmente para cada variable y para todo el sistema gramatical, también es posible comparar las cifras globales para los sistemas en cada momento. Debemos, por supuesto, observar que la difusión de los cambios estructurales ocurre según mecanismos propios de cada nivel de organización del sistema, y que una base de datos suficientemente amplia podrá permitir la identificación de correlaciones causales y cronológicas entre estos movimientos. En todos los estudios basados en corpora, el diseño de la base de datos es una función de la naturaleza de fenómenos analizados. Para los idiomas de que nos ocupamos en este trabajo, la variación morfosintáctica sólo puede ser adecuadamente delimitada en contextos frasales o más amplios, de la orden del párrafo. Naturalmente, la codificación de los datos incluirá tantas dimensiones cuantos sean los factores condicionadores seleccionados como hipotesis de trabajo (McEnery et al. 2006: 29ss). Además, es imperativo hacer test estadísticos de significancia para dar mayor seguridad a todos los procedimientos numericos que soportan el análisis (Mc Enery et al. 2006: 53ss; Biber et al. 2006: 275ss).

Ricart-Vayá , Alicia and María Alcantud-Diaz

Panel: 9. Usos específicos de la Lingüística de Corpus

USING COMPUTER- BASED CORPORA TO CREATE LEARNING MATERIALS FOR TOURISM (ESP)

The present article adopts a computerized frequency-driven approach in the analysis of frequently-used prepositions. Our purpose is to identify the errors made by first-year students of Tourism when writing essays. Thus, students were required to watch the film “The terminal” by Spielberg (2004) as a compulsory task which supplemented two of the units in their student’s book of English language (Walker and Harding 2006): “Airport departures” and “The airline industry”. The students were asked to write a film review of about 200 words. We decided to investigate the errors made when using the prepositions at, in and on. Our corpus was composed of 50 student’s essays, which we analyzed using WordSmith Tools 5 (Scott, 2010) both quantitatively and qualitatively. That is, we retrieved the prepositions analyzing their frequencies and concordances in order to look for non-native combinations. Our final aim was to create a series of exercises by using the occurrences of the prepositions as the main basis of our filling the gaps exercises. In this way, our students were provided with exercises based on their own errors when using prepositions in writing. These exercises created with the Exelearning programme for the design of learning objects, will be uploaded in the form of online activities for future students. As a general conclusion, we believe that corpus analysis could be an effective tool in order to

create tailor-made activities and teaching materials This research has been carried out within the frame of the Tur-i-Tic research team (Anglotic) from the University of Valencia.

Richa and Shahid Mushtaq Bhat

Panel: 7. Lingüística computacional basada en corpus

CASE SYNCRETISM IN URDU-HINDI: A CHALLENGE FOR NLP

This paper is an effort to bring into focus the key issue of Case Syncretism which is one of the challenges to the annotation of corpora in Indian languages both manually and automatically in terms of cognitive load to the annotator and computational complexity, respectively. The paper, based on the annotation of Hindi-Urdu corpora of 20K+ words, brings forth the issues of case syncretism. In this paper, Case Syncretism in Urdu-Hindi is explored from the perspective of corpus annotation, illustrating bottlenecks in the annotation process. This paper provides an optimal solution for manual tagging by offering linguistic rules specific to Urdu- Hindi. It also presents various disambiguating mini-algorithms for the automatic tagging. Finally, the analysis also shows that the residual issues can be handled at the level of argument structure as well as semantics. Thus, this paper supports a view that it is essential to annotate argument structure and semantic information for effective encoding of linguistic information and efficient POS-Tagging.

Roca-Varela, Mª Luisa

Panel: 8. Los córpora y la adquisición y enseñanza del lenguaje

CORPORA AS TOOLS AND RESOURCES FOR THE TEACHING OF ENGLISH VOCABULARY

Corpora are language databases which contain samples of real language use. These computerized databases are being increasingly used in both theoretical and applied linguistics with satisfactory results. This is true for both native and learner corpora (Granger, 1994; Palacios, 2005). In spite of this, the explicit application of corpora is relatively new within the field of applied linguistics, and the exploitation of corpora in the area of language teaching is not widespread. However, it has been proved that corpus-based language learning has positive effects and promotes learners' autonomy through data-driven learning (Johns, 1991; Leech, 1997). In this paper, I will first analyse how teachers can take advantage of these large language databases in EFL settings (Oghigian & Chujo, 2010) and how corpora can be useful resources for three basic areas of foreign language teaching: syllabus design, classroom materials and activities (Krieger, 2003). The second part of this paper will focus on the use of corpora for vocabulary teaching and learning. I will work on the information provided by native corpora (such as, the BNC or COCA) regarding the meaning and use of a particular lemma (collocations, colligations, semantic prosody). I will next show how useful it may be to compare and contrast native and learner corpus data to draw conclusions on what learners "know" about a L2 item and what they really "need to know". The ultimate goal of this study is to demonstrate the pedagogical usefulness of corpora for the teaching of vocabulary.

Rodríguez Arrizabalaga, Beatriz

Panel: 3. Estudios gramaticales basados en córpora

ON THE PRODUCTIVITY OF ENGLISH COGNATE OBJECTS. A CORPUS-BASED ANALYSIS

English cognate objects of the type a gruesome death in He died a gruesome death and an enigmatic smile in She smiled an enigmatic smile, for instance, have always being a matter of debate in linguistics

due to their controversial syntactico-semantic status (cf. Sweet 1891; Quirk et al. 1985; Rice 1987; Jones 1988; Massam 1990; Mittwoch 1997; Macfarland 1999; Pereltsvaig 1999; Felser and Wanner 2001; Kuno and Takami 2004 and Höche 2005, among others). As a consequence, the research carried out around this particular clause constituent has mainly focused on its syntactico-semantic behaviour, trying to look for an answer to the following problematic issues: (a) the own definition of the term 'cognate object'; (b) the syntactic function of cognate objects either as verbal arguments or adjuncts; (c) the syntactic verbal classes that are compatible with them; (d) the obligatory/optional patterns of modification they take; (e) the restrictions, if any, on the determiners that introduce them into discourse; (f) and the comparison, due to their semantic closeness, between cognate object structures and intransitive patterns with adverbial modification like He smiled in an enigmatic way, on the one hand, and light verb constructions of the type He had a gruesome death, on the other. In this debate, however, the pragmatic dimension underlying English cognate objects has almost gone unnoticed and, as a consequence, there are questions concerning their frequency, productivity, textual distribution and use that still remain unanswered. For this reason, and with the intention to shed some light on the real productivity of English cognates objects, I have carried out a thorough and exhaustive analysis in the British National Corpus of the four verbal classes that, according to Levin (1993), seem to be potentially compatible with cognate objects: namely, verbs of nonverbal expression, verbs of manner of speaking, waltz verbs and a fourth heterogeneous class that includes the verbs dream, fight, live, sing, sleep and think. The main objective of the present talk is to present the results of the aforementioned corpus-based study in order to prove, in agreement with Mittwoch (1997), that English cognate objects are "heavily restricted", as well as to account for the main reasons underlying their scarce productivity in the English language.

Rodríguez Arrizabalaga, Beatriz

Panel: 5. Corpus, estudios contrastivos y traducción

THE BIRTH OF A NEW RESULTATIVE CONSTRUCTION IN SPANISH

Whereas the English resultative construction of the type Peter hammered the metal flat has always been a common subject matter in English linguisitics due to its high level of occurrence in the English language (cf. Simpson 1983; Yamada 1987; Carrier and Randall 1992; Levin and Rappaport 1995; Goldberg 1995; Wechsler 1997 and Boas 2003, among other scholars), its Spanish counterpart, illustrated, for instance, in examples such as Coció un huevo duro and Cernió la arena fina, has not received enough linguistic attention because, contrary to what happens in English, its productivity has proved to be very restricted, appearing only in very specific contexts (cf. Bosque 1990; Demonte 1991; Mallén 1991; Demonte and Masullo 1999 and Rodríguez Arrizabalaga 2002). For some scholars, such a constructions is even said to be completely non-existent in Spanish (cf. McNulty 1988; Aske 1989 and Sanz 2000). Despite such a productivity imbalance being true, I will present in the present talk the results of an exhaustive analysis of the prepositional phrase hasta la muerte in the Corpus de Referencia del Español Actual (CREA), which reveal that, besides its intensifying and emphatic function illustrated, for instance, in Hay que animar al equipo hasta la muerte (CREA: 28), in the last few years such a phrase has developed a resultative attributive function, as can be seen, for example, in Las muejeres fueron torturadas hasta la muerte (CREA: 1) and Lo apedrearon hasta la muerte (CREA: 97), which has to be considered completely equivalent to that of the English resultative attributes dead and to death in sentences of the type of He shot the president dead or He shot the president to death. The extremely frequent and common appearance of the resultative attribute hasta la muerte in the media nowadays, as a natural consequence of the enormous impact on the mass-media of the negative social circumstances and conditions surrounding human beings (i.e., terrorism, wars, gender violence, etc.), on the one hand, and its proven presence in the CREA (Corpus de Referencia del Español Actual), on the other, are two clear reasons to state, thus, that this specific resultative construction, considered ungrammatical in Spanish some time ago for being a literal calque of the English resultative construction with dead or to death as attribute (i.e., He shot the president dead or He shot the president to death), if not having entered the language yet, is making its way directly into Spanish.

Rodríguez Martín, Gustavo Adolfo

Panel: 2. Discurso, análisis literario y corpus

TOPIC TRANSITION IN THE PLAYS OF BERNARD SHAW: SOME CORPUS-BASED REMARKS.

One of the most prominent features of Bernard Shaw's dramatic discourse is ideological density. His "theatre of ideas" turns the stage into a battleground for dialectical wit. Some of the key moments of these pithy conversations are topic-transition places. Many times, controversial questions have to be settled abruptly in the form of a summative statement; on other occasions, further discussion ensues due to disagreement between characters. Furthermore, disagreement can even be repaired by a call for restatement or correction. On the whole, Bernard Shaw resorts to a series of set phrases to mark these climactic points in his dialogues. The purpose of this paper is to analyse these set phrases, which have been identified and classified by using Wordsmith Tools ©.

Rodriguez-Puente, Paula

Panel: 1. Diseño, compilación y tipos de córpora

INTRODUCING THE CORPUS OF HISTORICAL ENGLISH LAW REPORTS: STRUCTURE AND COMPILATION TECHNIQUES

Since May 2009 the research group Variation Linguistic Change and Grammaticalization from the University of Santiago de Compostela has been working on the compilation of British English legal texts as a contribution to the version 3.2 of the larger multi-genre corpus ARCHER (A Representative Corpus of Historical English Registers). Taking as a point of departure the techniques employed for the selection and edition of texts for ARCHER, we have started the compilation of our own corpus of legal texts: The Corpus of Historical English Law Reports (CHELAR). This paper is intended to present the main structure and characteristics of the corpus, as well as the methodology used for its compilation. The new corpus will contain approximately half a million words and cover the years from about 1500 to 2000. The texts included are British English law reports: records of judicial decisions that are "cited by lawyers and judges for their use as precedent in subsequent cases" (EBO s.v. law report). The currently available corpora of legal English are mostly concerned with contemporary legal language (cf., e.g., the Cambridge Corpus of Legal English). Corpora of historical legal English include texts from Parliamentary acts, Royal orders and Privy Council's orders (cf. Anu Lehto's web page at http://www.helsinki.fi/varieng/people/varieng_lehto.html) and trial proceedings (cf. the Proceedings of the Old Bailey). Alternatively, the linguist interested in legal English from a diachronic perspective can resort to the legal texts included as part of larger diachronic corpora, such as the Helsinki Corpus (850-1710), The Lampeter Corpus (1640-1740) or the ARCHER Corpus (1650-1999). However, to the best of our knowledge, a computerized corpus of historical law reports has not yet been compiled. The Corpus of Historical English Law Reports will, therefore, constitute a new, useful resource for linguists with an interest in legal language, from both a synchronic and a diachronic perspective.

Romero-Trillo,Jesús, Silvia Riesco-Bernier, Karina Vidal, Belén Díez-Bedmar, Teresa Gerdes, Anna Gladkova, Elizabeth Lenn and Tíscar Espigares

Panel: 1. Diseño, compilación y tipos de córpora

CORPUS OF LANGUAGE AND NATURE (CLAN-PROJECT): THE REPRESENTATION OF LANDSCAPE UNIVERSALS IN LANGUAGE

The CLAN-Project intends to describe the cognitive representation of landscapes in speakers who live in different intercultural contexts, and their subsequent emotional responses via spoken language. The hypothesis of the study is that students of English as a second or foreign language or speakers of English as a Lingua Franca may not react in the same way to the perception of natural landscapes, as their responses might depend upon the emotional implications that a particular natural scenario may trigger based on their cultural backgrounds, as well as upon their command of the second or foreign language. For this purpose, the team has selected a series of descriptive variables to study the language produced by learners of English as a second or foreign language, and speakers of English as a Lingua Franca. The objective of the project is to sketch out an atlas of linguistic features that represents the different emotions manifested by landscape preferences on the basis of cultural self-identifications. As in previous studies carried out by our team, the corpus data will be used for linguistic analysis at different levels (phonological, lexical, syntactic, pragmatic, etc…) (cf. Romero-Trillo, 2008). Following an evidence-based and pragmatic approach, our project aims at the description of the cultural norms, values and social practices affecting the linguistic representation of landscape perception in a metalanguage that has equivalents in different languages (The Natural Semantic Metalanguage, Goddard and Wierzbicka, 2002) and can also be inscribed in the tradition of intercultural pragmatics and ethnopragmatics (Goddard, 2006). These approaches to the description of intercultural linguistic phenomena try to understand speech practices in their context with special attention to the culturally loaded words. It is important to mention that our approach is evidence-based and it relies on a pragmatic approach to (learner) corpora.

Roselló Verdeguer, Jorge

Panel: 8. Los córpora y la adquisición y enseñanza del lenguaje

EL USO DE LA PUNTUACIÓN EN TEXTOS DE ESTUDIANTES DE EDUCACIÓN SECUNDARIA

El trabajo del que a continuación damos cuenta es fruto de una experiencia didáctica realizada con alumnos de educación secundaria obligatoria y de bachillerato en un instituto público de Valencia con el objeto de mostrar, en un primer momento, el nivel de conocimiento de los signos de puntuación utilizados por los alumnos en textos escritos. A partir de esos datos, se diseñó una intervención pedagógica con el fin de darles a conocer su importancia y mejorar su uso. El corpus lingüístico que sirvió de base para realizar el análisis estaba constituido por un conjunto de producciones escritas por alumnos de ESO y de Bachillerato. Sobre estos textos, realizados al principio y al final de cada uno de los cursos que duró la experiencia, se realizó un estudio estadístico en el que las variables lingüísticas (dependientes) fueron los cinco signos de puntuación considerados básicos o de primer orden: el punto y aparte, el punto y seguido, el punto y coma, los dos puntos y la coma; mientras que las variables independientes guardaban relación tanto con factores sociológicos (edad, lengua habitual y nivel sociocultural) como con otros más específicos relacionados con nuestro trabajo: factores estilísticos (tipología textual), adscripción a un determinado grupo (experimental y de control), diferentes grados o niveles (ESO y bachillerato) y tiempo de realización (al principio y al final de cada uno de los dos cursos académicos). Una vez descritas las variables y codificadas sus diferentes variantes , se llevó a cabo un estudio estadístico de los 15.517 signos utilizados por los alumnos en sus escritos. Con la ayuda del programa informático SPSS, se realizaron análisis de frecuencias (tanto absolutas como relativas); tablas de contingencia, para observar el grado de asociación o incidencia entre las variables dependientes (signos de puntuación) y las variables independientes; pruebas como el ji cuadrado, con el fin de descartar la llamada hipótesis nula; análisis factoriales, para observar los agrupamientos producidos según el comportamiento conjunto de las variables, etc. Por último, para el tratamiento de la variación lingüística y la realización de análisis de regresión logística, se utilizó el programa Goldvarb 2001, que nos permitió ver cuáles eran las variables significativas en la correlación y el peso probabilístico de cada una de las variantes. Con este estudio estadístico pudimos ver los signos más utilizados por los estudiantes en sus escritos y los errores más frecuentes. Centrándonos en estos últimos, y aplicando métodos de la estadística inferencial, que –como se sabe– intenta extrapolar los datos a entidades mayores, también pudimos comprobar qué variables resultaban estadísticamente significativas en cada uno de los signos y qué factores estaban contribuyendo a que se produjera el error. Todos estos datos

estadísticos extraídos del corpus nos permitieron realizar unas propuestas didácticas encaminadas a mejorar el uso de los signos de puntuación en los textos escritos por los estudiantes.

Rossini, Rema, Fabio Tamburini and Andrea Zaninello

Panel: 9. Usos específicos de la Lingüística de Corpus

EXPLOITING CORPUS EVIDENCE FOR AUTOMATIC SENSE INDUCTION

In this paper we intend to explore how statistical analysis and corpus evidence can contribute to sense disambiguation in non-annotated text. We focus on collocations as a source of surface evidence automatically extracted from corpora through positional and association-based procedures following probabilistic criteria. Our basic assumption is that most characteristic collocates of a (potentially polysemic) word are a good indicator of its meanings and that co-occurrence frequencies can be used to discriminate between different senses, in line with the Firthian tradition and the classical Harrissian distributional hypothesis. Our paper is organized as follows: firstly, we present a brief description of CORIS, the 120-million-word reference corpus of written Italian used in our study, composed of common, authentic texts chosen by virtue of their representativeness of modern Italian (cf. Rossini Favretti, Tamburini & De Santis 2002). Secondly, we describe the analysis tools exploited in our research. Thirdly, we present some case studies focusing on highly polysemic words in Italian. Collocation sets for the node are created through an automatic, iterated process of collocation analysis based on association measures and recursively applied to the collocates. The results are represented as co-occurrence graphs (cf. Heyer et al. 2001). This representation, formerly exploited to modulate register variation (cf. Rossini & Tamburini 2009), allows one to single out clusters of collocates connected at different strengths, and thus define different meaning areas providing a visualisation of polysemy through a representation of the collocates' distribution in a vectorial semantic space. As a matter of example, we analyse the collocates of the node "calcio" (meanings football, calcium, kick…) which are organised around two main axes, corresponding to the two main senses of the word:

1) Chemistry (meaning: 'calcium')

Pattern 1.a: NOUN+PRE+NOUN* (e.g. carbonato di calcio) - Asymmetric relation: node modifies collocate

Pattern 1.b: NOUN+COORD+NOUN* (e.g. calcio, potassio e magnesio…) – Symmetric relation: node and collocate are co-hyponyms

2) Sport (meaning: 'football')

Pattern 2.a: Cranberry (e.g. 'Quelli che il calcio') – Arbitrary node-collocate relation

Pattern 2.b: NOUN+PRE+NOUN* (e.g. squadra di calcio, campo da calcio); Symmetric relation: node modifies collocate

Pattern 2.c NOUN+COORD+NOUN* (e.g. calcio e basket); Symmetric relation: node and collocate are co-hyponyms

Pattern 2.d NOUN+ADJ (e.g. calcio italiano) Asymmetric relation: collocate modifies node

We conclude that the main senses of the node can be identified fairly accurately by the clustering procedure. However, the kind of relationship between the collocate and the node (co-hyponymy, kind-of relation etc.) are consistent with and can only be identified by an analysis of the linguistic structures they feature in, making an integration of the two procedures desirable. As a suggestion for future work, we believe this procedure may be applied to multiword units to measure their level of opaqueness comparing the collocates of the head with the collocates of the MWU taken as a whole. Moreover, in order to study the evolution of a word's senses across time, this procedure may be expanded in a historical dimension by applying it to diachronic corpora such as DiaCORIS, a representative and

balanced collection of Italian written language ranging from the National Unification (1861) to the end of the Second World War (1945) (cf. Onelli et al. 2006).

Ruano-Garcia, Javier

Panel: 6. Corpus y variación lingüística

THE WORLD HAS GOT SOME HINT OF HER COUNTRY SPEECH: ON THE ENREGISTERMENT OF THE 'NORTHERN DIALECT'

Recent research in sociolinguistics and dialectology has introduced the concept enregisterment to refer to the process whereby certain linguistic features become associated with a particular place and specific sociocultural values (see Agha 2005, 2007; Beal 2009a; Johnstone, Andrus and Danielson 2006; Remlinger 2009; among others). Agha (2003: 231) defines it as "the processes through which a linguistic repertoire becomes differentiable within a language as a socially recognized register of forms". Some studies exemplifying it have shown that enregisterment occurs through a series of discursive practices. For example, Beal (2010: 94-95) asserts that "speakers/writers may take part in the process of enregisterment via such practices as dialect writing, the compilation of dialect dictionaries and, more recently, websites dealing with issues of dialect and local identity" (see further Beal 2009b). This paper places literary renditions of northern English into the context of enregisterment. It investigates the repertoire of forms which have commonly been identified as northern and have, thus, contributed to the enregisterment of the 'northern dialect'. For this purpose, I shall undertake a corpus-based analysis of literary texts included in the Salamanca Corpus, laying emphasis on early modern material. My aim is threefold. Firstly, to identify the most recurrent traits of these representations, and the sociocultural values they index. Secondly, to ascertain if the set of forms depicted in the early modern literary discourse was maintained across time by surveying selected corpus material from the late modern period. Thirdly, to show that dialect writing, though much neglected for linguistic research, gives insights into language variation and attitudes. In fact, these texts are inextricable from the historico-linguistic context in which they were produced, and from the attitude(s) towards the 'other' English which they reproduce (see Dawson and Larrivée 2010, for example).

Sánchez Aquilino, Cantos Pascual and Criado-Sánchez Raquel

Panel: 8. Los córpora y la adquisición y enseñanza del lenguaje

CORPORA-BASED FREQUENCY LISTS, READABILITY INDEX AND ELT TEXTBOOKS

Vocabulary frequency lists for the elaboration of FLT/L (Foreign Language Teaching/Learning) materials were already used in the first half of the 20th century (Thorndike, 1924, 1944; García Hoz, 1953; West, 1953). No doubt, vocabulary lists based on modern corpora are more reliable, valid and, above all, 'real' (Kucera & Brown, 1967; Sinclair, 1987; Leech & Al., 2001). Corpus-based FLT/L materials have become a must (Johns, 1994; Sinclair, 1996). The underlying rationale is that the most frequently used words in a language are also likely to be the most useful for communicative purposes. Hence, the learning of those very common words turns into one of the most important priorities in language teaching and learning, since effectiveness in communication, vocabulary frequency and communicative potential are intrinsically interwoven. The emphasis on vocabulary control and grammar has decayed significantly in the Communicative Approach, and the focus is placed instead in the communicative functions of language. In spite of this bias towards content and meaning, the popularity of corpora in linguistic research has maintained the interest for frequency lists and their importance in language teaching. In addition to that, many studies refer explicitly to the first 1,000, 2,000, 3,000, etc., words and the role they play in establishing effective communication (Nation, 1990; Diller, 1978; Gildea, 1987; Laufer, and Nation, 1995; Waring, 1997; Zechmeister et al., 1995). The popular appeal of ambiguous and biased slogans such as 'Learn the first 1,000 words of English' also contributes to increasing the importance of learning 'the most frequent words of the language'. This paper addresses the issue of whether the

teaching materials have adapted or not, and to what extent, to the underlying beliefs and convictions regarding vocabulary teaching in connection to frequency lists. We shall investigate here whether claims and expectations on frequency lists and their role in teaching are really reflected in textbooks. Our research is based on the vocabulary analysis of a widely used textbook in the context of Spanish Secondary Education: English in Mind – Student's Book 2 (Puchta & Stranks 2005). For this investigation, we compiled an ad hoc corpus with the whole textbook content. Our aim is (i) to quantify and typify new vocabulary and new vocabulary rate per unit in the textbook, (ii) to correlate the textbook vocabulary and frequency list against the BNC-based frequency list and ranges (Nation, 2001), and finally (iii) to determine the readability index (text reading difficulty) of the texts as found in this course book. To achieve these goals, we first extracted the tokens and types present in the textbook, then we systematized the data obtained and identified the new words per unit; we later contrasted all the vocabulary items in the textbook against the BNC frequency list, in order to discover whether both lists matched or not and to which extent. Finally, we calculated the readability index –applying the ARI - Automated Readability Index, based on word and sentence length– and related it to the size of the new vocabulary in the textbook. The results of our analysis provide (i) a reliable picture on how a contemporary and widely used textbook adapts to the claims regarding the relevance and function of frequency lists in FLT/L materials, and ii) a reliable readability index (ARI) to determine the difficulty of the materials built with the vocabulary analyzed.

Sánchez Cárdenas, Beatriz and Pamela Faber Benítez

Panel: 4. Lexicología y lexicografía basadas en córpora

LA PROTOTIPICIDAD DE LOS ARGUMENTOS VERBALES COMO FACTOR DELIMITADOR DE LA ESTRUCTURA JERÁRQUICA DE UN DOMINIO LÉXICO

La producción discursiva responde a un equilibrio entre reglas gramaticales y preferencias de selección léxica (Bosque 2004). Las segundas son claves para determinar las dependencias jerárquicas entre los verbos de un mismo dominio léxico. Mediante el análisis de corpus establecemos una clasificación de la frecuencia de uso de los distintos tipos de argumentos de un grupo de verbos. Estos datos conducen a la estructuración jerárquica del dominio en cuestión. En definitiva, demostramos cómo el análisis de corpus puede ayudar a dirimir las relaciones de interdependencia jerárquica entre los verbos y cuál puede ser su aplicación a la lexicografía. Al hablar de jerarquías del léxico, no se pueden pasar por alto los trabajos en lingüística cognitiva sobre el léxico nominal (Rosch 1975 ; Lakoff 1987; Kleiber 1990). Se ha prestado, sin embargo, menos atención al léxico verbal, tradicionalmente dejado de lado por considerarse que los verbos no responden a la canónica estructura en tres niveles (superordinado, básico y subordinado: fruta/manzana/golden) (Rosch 1975). Si bien es cierto que esta jerarquía tripartita de los sustantivos no se da en el ámbito verbal, las relaciones de dependencia semántica existen igualmente entre los verbos y tienen, como veremos, una importancia primordial (Faber & Mairal 1999). Partimos de la premisa de que cada verbo lleva asociados predicados prototípicos. Si tomamos el ejemplo del verbo comprar, veremos que "María compra" es más prototípico que "el cangrejo compra" . De modo que el grado de prototipicidad de los argumentos verbales es fundamental para determinar su significado (Fellbaum 1990, Kleiber 1990: 129). La metodología implementada se basa en la observación de que los verbos situados en las posiciones jerárquicas superiores son los que admiten un mayor número de argumentos prototípicos. En contrapartida, a medida que descendemos en la escala de dependencia, los verbos se vuelven más restrictivos, admitiendo cada vez un menor número de categorías semánticas en sus argumentos. El análisis de corpus lleva a la clasificación de las categorías más prototípicas de los argumentos de cada verbo, basándonos en la tipología nominal de Flaux & Van Velde (2000). Las unidades léxicas que sirven de ejemplo a nuestro análisis son los verbos de cuantificación en francés, un dominio poco estudiado hasta ahora. El corpus, analizado con la aplicación WordSmith tools 5.0, ha sido constituido a partir de: (a) la base de datos "Frantext"; (b) las actas del Parlamento Europeo recopiladas en "Corpusye"; (c) el corpus periodístico de Chambers & Rostand (Universidad de Oxford); (d) la base de datos "Wortschatz" (Universidad de Leipzig); y, finalmente, (e) una cuidada selección de documentos extraídos de buscadores en internet. Del análisis de corpus se desprenden datos estadísticos en cuanto a la frecuencia de uso de cada categoría nominal, lo que permite conocer el grado de restricción de cada verbo con respecto a sus argumentos. Esto arroja indicios inequívocos sobre la organización jerárquica interna del dominio en cuestión. Esta metodología

permite dar el salto a nuevos tipos de diccionarios onomasiológicos que den cuenta de las redes semánticas entre palabras. La investigación en lexicografía no debe eludir la estructuración jerárquica del léxico verbal si pretende dar cuenta de la organización cognitiva del léxico.

Sánchez-García, Pilar

Panel: 6. Corpus y variación lingüística

THE WESTMORELAND DIALECT IN THREE DIALOGUES (1790): THE CONTRIBUTION OF ANN WHEELER'S DIALOGUES TO JOSEPH WRIGHT'S THE ENGLISH DIALECT DICTIONARY.

Joseph Wright's monumental work English Dialect Dictionary (1898-1905), the most comprehensive dialect piece hitherto compiled, is much indebted to thousand of works, both literary and non-literary pieces, as he himself acknowledges in the preface to his work 'upwards of three thousand dialect glossaries and works containing dialect words have been read and excerpted for the purposes of the Dictionary' (vi). As it is well known, the volume of works corresponding to the counties of Lancashire and Yorkshire exceeds the number of works of the other four northern dialects for evident reasons. There are emblematic pieces corresponding to Lancashire and Yorkshire thoroughly analysed and studied while other important pieces of the many other dialects remain almost unnoticed. If there is an emblematic work representing the dialect of Westmoreland that is Ann Wheeler's The Westmoreland Dialect in Three Familiar Dialogues (1790). Commentaries to this work such as that appeared in Russell Smith's list of "interesting books" included in his Bibliographical Lists (1839): "The philologist will find numerous examples of words and phrases which are obsolete in the general language of England, or which are peculiar to Westmoreland and Cumberland from time immemorial" (7) have made us consider the importance of undertaking and in-depth analysis of this dialogues. This paper tries to evaluate the contribution made to Wright's English Dialect Dictionary by Wheeler's dialogues, considering not only the first edition (1790) but also later editions of this work (1802) and (1840) to which there are important additions. This undertaking has been much more feasible thanks to the digitised version of Wright's English Dialect Dictionary being prepared by the research team at the University of Innsbruck (Markus 2007, 2009, Markus & Heuberger 2007). Our aim is twofold. Firstly, to ascertain the entries from Wheelers's dialogues included in Wright's masterpiece and to analyse the treatment given to this information. Secondly, to contribute to a better knowledge of one of the northern dialects which traditionally has received poor attention, the Westmoreland dialect.

Sanmartín Sáez, Julia and Nuria Edo Marzá

Panel: 4. Lexicología y lexicografía basadas en córpora

ANÁLISIS DEL CONCEPTO 'HABITACIÓN' EN UN CORPUS BILINGÜE ESPAÑOL-INGLÉS DE PÁGINAS ELECTRÓNICAS DE PROMOCIÓN HOTELERA

El presente artículo, concebido en esta fase como ensayo piloto, muestra uno de los tipos de estudio – concretamente de tipo léxico– que pretendemos llevar a cabo en el proyecto titulado Análisis léxico y discursivo de corpus paralelos (Español-Inglés-Francés) en páginas electrónicas de promoción turística. Este artículo tiene, pues, como objetivo principal el análisis contextualizado de las unidades léxicas que correspondan al concepto 'habitación' a través de los datos obtenidos mediante la creación y explotación de un corpus ad hoc en dos lenguas de trabajo. El corpus de estudio está formado por un conjunto de páginas electrónicas de hoteles de cinco y cuatro estrellas de España, y Chile (corpus de versión original en español y su traducción en inglés) e Inglaterra y EE.UU. (corpus de versión original en inglés y su traducción en español). A partir de dicho corpus procederemos al establecimiento de las posibles unidades léxicas especializadas que cubran el concepto 'habitación' (y sus tipos) en versiones traducidas y originales de las dos lenguas de trabajo. Para el establecimiento de estas unidades, tomamos como punto de referencia inicial la descripción y la propuesta de términos señaladas en las normativas internacionales de estandarización en materia turística (Norma ISO 18513 Tourism Services -

Hotels and other types of tourism accommodation - Terminology) y en las legislaciones de las comunidades autónomas (para el caso del español). Este punto de referencia nos facilita la búsqueda en el corpus de unas unidades léxicas concretas y el análisis de sus colocaciones. De este modo, el proceder estrictamente lexicológico (del lexema a su significado) se combina con el terminológico (del concepto al término), tal y como se postula desde perspectivas terminológicas comunicativas a (Cabré 1993). Además, tras la extracción de las unidades léxicas, se deberá distinguir qué colocaciones constituyen realmente unidades designativas en el discurso especializado objeto de estudio (habitación adaptada) y cuáles son meramente sintagmas recurrentes (habitación con vistas). De este modo, determinaremos las unidades léxicas que configuran el concepto de 'habitación' y sus tipos en dos modalidades geográficas del español y del inglés, y, además, podremos comparar las versiones originales de estas unidades con sus traducciones.

Santaemilia Ruiz, José and Sergio Maruenda-Bataller

Panel: 2. Discurso, análisis literario y corpus

BUILDING A COMPARABLE CORPUS (ENGLISH-SPANISH) OF NEWSPAPER ARTICLES ON GENDER AND SEXUAL (IN)EQUALITY (GENTEXT): PRESENT AND FUTURE APPLICATIONS IN THE ANALYSIS OF SOCIO-IDEOLOGICAL DISCOURSES

Over the last few years a number of legal measures have been adopted recently both in Spain and in the UK –e.g. the Civil Partnership Act 2004 or the Domestic Violence, Crime and Victims Act 2004 in UK, or the new Spanish legislation on abortion, gender-based violence or homosexual marriages. These measures, along with the growing recognition of social and sexual rights in Western Europe, have sparked a heated debate within both Spanish and British societies. These debates are reproduced, generated, amplified, diminished, perverted or exploited by mass media, political parties or religious institutions, with a view to demanding or imposing either respect or neglect for the very minorities to which these legal measure are addressed. As part of the work of the research group GENTEXT , we have built a 4.5 million-word, comparable (Spanish-English) , highly-specialised corpus (GENTEXT-N) which serves to analyse, document and offer insights into the complex socio-ideological debates behind people's attitudes and values, into the discursive attempts to exercise power and to impose political and religious positions, and so on. It is, in short, an invaluable source of material to document the steps our societies are making towards sexual equality. We believe that a combination of qualitative and quantitative analyses (as advocated, among others, by Baker & McEnery 2005, Baker et al 2008, Caldas-Coulthard 2010) is essential if we wish to grasp both the linguistic and the ideological underpinnings of the heterogeneous texts we are investigating. Thus, our analyses integrate, on the one hand, critical discourse analysis and lexical semantics/pragmatics and, on the other, Corpus Linguistics techniques to fully exploit the potentialities of both approaches, thus trying to avoid the oversimplification of ideological bias. We will start with a statistical keyword analysis, using WordSmith Tools, in order to have a reliable list of recurrent words in the field. As argued by Baker (2006), keywords are not only neutral or statistical lists of words, but rather privileged rhetorical devices used to implant common sense in our ways of thinking. This analysis will provide information on the ideological implications of keywords, as well as on the ideas these keywords cluster around (naming strategies, 'sensitive' relationships between minority group members, social and sexual implications, identities and self-presentation …). Attention to context will be paramount. Apart from the initial focus on keyword in these gender-sensitive texts, we also examine the potentialities of collocations and semantic/discourse prosodies for our research (Louw 1993, Stubbs 2001). As for the former, it will be revealing to document the collocations certain keywords (e.g. homosexual, abortion or violence) give rise to. As for the latter, semantic/discourse prosodies help transcend the collocational or even sentential scope to reveal discursive patterns and, consequently, to trace evaluative relations (with participants in discourse) in terms of ideological standpoint (see Martin & White 2005). These constitute a network or constellation of semantic concepts which contribute to shaping and (de)legitimising citizens' discourses and rhetorical frameworks within communities of practice.

Santos Moreira, Adonay Custódia

Panel: 1. Diseño, compilación y tipos de córpora

TURIGAL: COMPILATION OF A PARALLEL CORPUS FOR BILINGUAL TERMINOLOGY EXTRACTION

These last few years have witnessed an increase in research involving the compilation of large quantities of texts and their respective translations, as well as the development of techniques for processing those bilingual term banks (Bowker & Pearson, 2002; Biber et al., 2004; McEnery & Wilson, 2004). The present study is an example of such research as it uses a Portuguese-English unidirectional parallel corpus as a starting point for the retrieval of terminology. The main goal of this research is to exploit one of the possibilities offered by parallel corpora: the compilation of bilingual term banks. Turigal, a parallel corpus of tourist advertising material, has been devised to support the creation of a bilingual term bank on tourism. The corpus consists of texts – printed brochures and websites – in Portuguese and their translations into English, all of which were sourced from Portuguese Tourism Regions, Regional Tourism Boards and Regional Tourism Promotion Agencies, and stored as plain text. For the moment, it contains 1,285,764 words (632,193 words in Portuguese and 653,571 in English) and it is included in the Linguistic Corpus of the University of Vigo (Gómez Guinonart, 2003) and available for free consultation at http://sli.uvigo.es/CLUVI. Turigal is considered to be sufficiently representative of all bilingual (Portuguese-English) promotional materials published and distributed by the official entities responsible for the internal and external tourism promotion of Portugal in 2007, the year the texts were collected. First, we describe the methodology used in the compilation of Turigal. Then, we discuss Pearson's (1998) set of criteria for corpus design and text selection – namely size, text origin, author, factuality, technicality, audience, intended outcome, setting and topic – which has been considered when compiling our corpus. Finally, we present the alignment and tagging of Turigal. The programme TRANS Suite 2000 Align (Cypresoft, 2000) has been used to align the texts. All the aligned parallel texts are stored in TMX format and three translation strategies – omission, addition and reordering – have been encoded.

Del Saz Rubio, M. Milagros

Panel: 2. Discurso, análisis literario y corpus

AN APPROACH TO NATIVE AND NON-NATIVE WRITERS' USE OF INTERACTIONAL METADISCOURSAL FEATURES IN SCIENTIFIC ABSTRACTS IN ENGLISH WITHIN THE FIELD OF AGRICULTURAL SCIENCES

The relevance of academic writing is nowadays more than justified as demonstrated by the large body of research in this area. Authors such as Berkenkotter, Huckin & Ackerman (1991) have brought to attention the importance of mastering a *specialized literacy*, especially for students or researchers entering the academic disciplines. This literacy can be defined as the ability to make use of the discipline-specific rhetorical and linguistic conventions in order to fulfill the purpose as writers. Mastering academic writing thus involves an awareness of the existence and structure of specific genres, as a key element for acculturation and success. Therefore, to engage in the writing of a genre such as the research article (Ra), inevitably calls for awareness of its specific conventions, as well as of the role of the writer and the purpose of the writing task. This situation can be certainly more complex for researchers who need to write and publish their research in an L2, since mastering the grammar, lexicon or syntax is not enough to guarantee them communicative competence. Taking all this into consideration, the main aim of this paper is to assess whether there is intercultural variation in the rhetorical preferences of native English and Spanish-speaking researchers when writing research articles abstracts in English within the field of Agricultural sciences. To do so, a total of 30 articles, 15 written by native English speakers (NES) and 15 by non-native English speakers (NNES) are analysed and a quantitative and qualitative analysis of the interactional metadiscoursal features they employ, as developed by Hyland (2005), is carried out using WordSmith Tools 4.0. By focusing on the use that these writers make of interactional devices in the different sections of the abstract, the use of hedges and boosters, engagement and attitude markers and self-mentions will be looked into as they are devices traditionally employed by writers to involve the reader in the text and thus explicitly build a relationship with the scientific audience. Results will also aid to determine if it is possible to talk of the existence of a conventional international culture in the genre of research articles within the field of Agricultural

sciences, or if, on the contrary, the two groups of writers tend to impose the writing conventions of their L1(s) in their writing of abstracts for scientific articles in English and thus, in their use of metadiscoursal features. Finally, the results obtained here can have implications for the teaching of academic writing to non-native speakers of English. As such, they will be taken as a starting point for the design and elaboration of meaningful writing activities aimed at raising awareness of the conventions and expectations which operate in the genre of the research article in English in the field of Agricultural Sciences.

Schneider, Gerold and Fabio Rinaldi

Panel: 6. Corpus y variación lingüística

A DATA-DRIVEN APPROACH TO ALTERNATIONS BASED ON PROTEIN-PROTEIN INTERACTIONS

Syntactic alternations, for example the dative shift, are well researched. There are investigations using large amounts of quantitative data and statistical techniques (e.g. Bresnan and Nikitina 2009). Recently, it has been suggested that traditional concepts of alternations are a heritage from generative syntax, that most decisions which speakers take are more complex than binary choices (e.g. Arppe et al. 2011) and there are complex interdependencies and combination options (e.g. Fillmore 2003). Multifactorial approaches and, as speakers choose among a wide range of grammatical forms, a large inventory of syntactic patterns need to be considered to supplement current approaches. We use the term semantic alternation, broadly referring to the many different ways in which a relation between entities, conveying broadly the same truth-functional value can be expressed. We use a clearly defined and well-resourced domain, biomedical research texts, for a corpus--driven approach. As entities we use proteins, and as relations we use interactions between them, using data from large applied text mining challenges (e.g. Leitner 2010). The following sentences all convey the same core relation:

We confirm binding of MEA to FIE

In our experiments MEA amino acids were able to bind to FIE

FIE-binding by MEA amino acids has been observed

The amino acids of MEA are sufficient to bind with FIE

We discuss first an approach using a finite inventory of manually designed syntactic patterns, second a corpus-based semi-automatic approach and third a machine-learning language model. The machine-learning approach learns the probability that a certain syntactic configuration expresses a relevant interaction of given event types from an annotated corpus. For each event, the inventory and probabilities of configurations define the envelope of application and its multitude of forms. A configuration consists of dependency relations and lexical chains, which use semantic information to overcome sparse data problems. As it has been pointed out that predictive models are particularly accurate for expressing complex, multifactorial phenomena (e.g. Tse 2003), we thus present and evaluate a predictive probability model for semantic alternations in the domain and also discuss its relevance for other domains.

Seghiri, Miriam

Panel: 1. Diseño, compilación y tipos de córpora

COMBITUR: ASPECTOS DE DISEÑO, COMPILACIÓN Y REPRESENTATIVIDAD DE UN CORPUS DE CONDICIONES GENERALES DE VIAJE COMBINADO

En este trabajo se presenta una metodología protocolizada para la compilación de corpus virtuales, que se ilustrará mediante la creación de un corpus de condiciones generales de contratos de viaje combinado en lengua inglesa, denominado CombiTur, creado únicamente a partir de recursos electrónicos en red. Esta metodología se estructurará en dos fases: en primer lugar, criterios de diseño,

y, en segundo lugar, un protocolo de compilación dividido cuatro pasos, que garantizará la representatividad cualitativa de la colección. Una vez compilado el corpus CombiTur, se hará uso de la aplicación informática ReCor con objeto de determinar la representatividad cuantitativa de la muestra. De este modo, el corpus CombiTur estaría en condiciones de ser utilizado para la realización de estudios lingüísticos y traductológicos sobre la contratación de viajes combinados en lengua inglesa. La elección de las condiciones generales de viaje combinado se justifica por la enorme demanda de este tipo de traducciones, tanto directa como inversa, en nuestro país (cfr. ACT, 2005), ya que la industria turística española representa uno de los pilares fundamentales de nuestra economía, pues tal y como apuntan Alcaraz Varó et al. (2006) «El turismo de masas es uno de los fenómenos más novedosos desde la segunda mitad del siglo xx. En España es la industria número uno, la fuente principal generadora de riqueza y de puestos de trabajo».

Skorczynska, Hanna

Panel: 2. Discurso, análisis literario y corpus

METAPHOR IDENTIFICATION IN CORPORA: THE CASE OF 'AS' IN A BUSINESS PERIODICAL CORPUS

Metaphor signals, also called metaphorical markers (Goatly, 1997), tuning devices (Cameron & Deignan, 2003) and flagging expressions (Steen, 2007), are words and phrases that anticipate metaphors in discourse and are meant to cue the reader/listener into the metaphorical rather than the literal interpretation of an expression. Metaphor signals also provide a direct access to metaphorical material in large corpora if concordancing techniques are used. Goatly (1997), as well as Wallington et al. (2003) have proposed the listings of possible metaphor signals. Their use in electronic queries of corpora may be an alternative to the troublesome manual searches for metaphors, and other corpus techniques, such as Charteris-Black's (2004) use of key metaphors. The studies of metaphor signals in different types of discourse and corpora (Skorczynska & Piqué, 2005; Wallington et al., 2003) have shown that only a small percentage of all metaphors used are signaled. In spite of that, metaphor signals can still be used as a complementary metaphor identification procedure, given that no reliable metaphor identification computer tool has been designed to date. The use of metaphor signals in metaphor identification methods still needs to be refined through the analysis of language data extracted from corpora. The potential metaphor signals should be evaluated with regard to the probability with which they fulfill this function. They also need to be examined in their co-text to identify larger phraseological chunks that might anticipate metaphorical expressions in discourse. Both the probability and the phraseology might vary in different discourse types and corpora. In response to these needs, this study looked into the use of 'as' as a metaphor signal in a corpus of business periodicals. A previous study (Skorczynska & Piqué, 2005) had revealed that 'as' was one of the most frequent words signaling a metaphor in this corpus. In the present study a corpus of around 600,000 words was electronically queried for 'as'. Of the 4,772 occurrences, which were manually analyzed, only 260 (5.5%) were used to signal a metaphor. The co-text of these occurrences was further analyzed in order to determine possible metaphor signaling phraseological patterns. One of the patterns identified was the combination of a verb with 'as'. The following combinations were, therefore, further examined: 'view as', 'refer to as', 'describe as', 'look as', 'act as', 'perceive as', 'think of as', 'see as', 'know as', 'use as', 'call as'. The comparison of metaphor signaling uses of these word combinations with their non-signaling uses showed that some of them are more probable metaphor signals than others. For instance, 'view as' was found to be the most reliable metaphor signal that registered 75% of metaphor signaling uses. The least probable metaphor signal was 'call as' with only 18% metaphor signaling occurrences. The results obtained suggest that the use of 'as' combined with other lexical items as the search words in corpus electronic queries might be a more efficient metaphor identification technique than 'as' on its own.

Soler-Monreal, Carmen and Luz Gil-Salom

Panel: 6. Corpus y variación lingüística

LITERATURE REVIEWS IN ENGLISH AND SPANISH PHD THESES: A CROSS-LANGUAGE STUDY

Research on the Literature Review (LR) chapter of doctoral theses has been carried out on theses produced by native English speaking students (Ridley, 2000; Kwan, 2006; Thompson, 2005, 2009). However, to our knowledge there have been no contrastive studies based on LR chapters in theses written in English and in Spanish. Reviews in general entail critical evaluations which may involve face threatening acts (Brown & Levinson, 1987). The LR of a doctoral thesis both evaluates others' research and is evaluated by the examiners, a distinguishing feature of the thesis social context which differentiates it from other 'more public' review genres, such as book reviews or back-cover blurbs (Hyland & Diani, 2009). This makes it necessary to maintain appropriate relations between the writer and the academic community through politeness strategies that aim at saving three faces: the writer's, the examiners' and the reviewed authors'. Doctoral candidates must submit their research for assessment and need to present their claims and show their knowledge in conformity to the norms of the academic environment. Citation practice provides justification for arguments and allows a writer to indicate a rhetorical gap for her/his research and adopt a tone of authority. Claims must be supported with evidence, and writers must demonstrate an understanding of approaches and knowledge in their fields of specialisation, in order to persuade the examiners that the thesis is worthy of the award of a doctorate (Thompson, 2005). Candidates also need to keep the adequate interpersonal relationship with the immediate audience (the examiners). They also need to evaluate the previous research in an area of study and to be respectful with previous claims from authorities in the disciplines. In this context of social interaction, politeness strategies should be taken into consideration so as to mitigate the strength of their arguments. This paper investigates contrastively how interactional resources and, in particular reporting verbs, are deployed in the LR chapters of PhD theses. It analyses a comparable corpus of 20 LRs -10 in English and 10 in Spanish- written by native speakers, within a single applied discipline: computing. It focuses on uses of reporting structures realised through integral and non-integral citations of other texts (Hyland, 1999). The research design is based on previous taxonomies of reporting verbs proposed by Thompson & Ye (1991) and Hyland (1999), and classified according to the type of activity referred to, under two categories: denotative, e.g. 'find', 'state' (in English LRs) and 'demostrar', 'analizar' (in Spanish LRs), and evaluative, e.g. 'suggest', 'recommend' (in English LRs) and 'proponer' 'asumir' (in Spanish LRs). Using a combination of both quantitative and qualitative data we will determine if there is some variation in the way English and Spanish doctoral candidates adopt a stance to their reviewed authors. The pedagogical implications of this study will contribute to an understanding of interpersonal relations in two different cultural and linguistic backgrounds, and will help novice academic writers interact with their intended readers successfully.

Keith Stuart

Panel: 2. Discurso, análisis literario y corpus

A CORPUS ANALYSIS OF RHETORICAL STRATEGIES IN THE DISCOURSE OF CHOMSKY

This paper explores the rhetorical strategies used by Chomsky in two of his most important books (Syntactic Structures, 1957 & Aspects of a Theory of Syntax, 1965). It continues and widens the research carried out by Hoey (2001) who has analysed Chomsky's rhetorical strategies but limited his study to just two passages of Chomsky's writings. One of the claims that we shall be making is that Chomsky is an expert in wrapping propositions in the form of interpersonal metaphors so as to appear objective and factual. In interpersonal metaphors of modality, the grammatical variation which occurs is based on the logico-semantic relationship of projection (Halliday, 1994: 354). In other words, Chomsky construes propositions as projections and encodes the "objectivity" in a projecting clause. I have found 264 clause complexes of this type in Aspects of a Theory of Syntax and 248 in Syntactic Structures. Some examples are given of the way Chomsky dissimulates that he is expressing an opinion through the use of the logico-semantic relationship of projection.

it seems quite clear that no theory of linguistic structure… (Who is it clear to?)

it is unquestionable that opposition to mixing levels, … (Who thinks it is unquestionable?)

It is quite true that the higher levels of linguistic description... (Who is it true to?)

The paper will not limit itself to these structures but analyzes a range of interpersonal meanings and their lexico-grammatical realizations. It will also make reference to a recent paper by Pullum (forthcoming, 2011) on the mathematical foundations of Syntactic Structures and suggest some reasons why Chomsky dresses up his texts in a very persuasive form of language. These reasons seem to be principally issues to do with the academic and historical context in which these two important texts were produced.

Therón, Roberto

Panel: 4. Lexicología y lexicografía basadas en córpora

ANÁLITICA VISUAL: UN NUEVO ENFOQUE EN LA LINGÜÍSTICA DE CORPUS PARA EL NUEVO DICCIONARIO HISTÓRICO DEL ESPAÑOL

Actualmente, la Lingüística, con ayuda tecnológica es capaz de generar cantidades de datos mucho más grandes de las que el lexicógrafo puede abarcar (éste es el caso del NDHE y su corpus, CDH [Pinillos, 2008]). Esto hace extremadamente difícil adquirir una "idea global" del mismo. El problema se hace todavía más complejo con la naturaleza geolocalizada y diacrónica de los datos. Tal cantidad de datos necesita nuevas herramientas basadas en software, y su complejidad requiere que se tenga en especial consideración su representación visual si se quiere destacar características según su importancia, descubrir relaciones en los datos y mostrar fenómenos geográficamente localizados ocurridos a lo largo de la historia. Así, este trabajo presenta los primeros logros alcanzados en dos aspectos: 1) Corpus del NHDE (CDH) Inteligente. Soluciones para los procesos de explotación, análisis y validación dirigidos por expertos, de forma automática e inteligente, en el CDH. 2) Mapa de diccionarios. Soluciones visuales interactivas que facilitan los trabajos linguísticos relacionados con la evolución temporal en el NDHE.

Toledo Báez, María Cristina

Panel: 5. Corpus, estudios contrastivos y traducción

TRANSLATING RESEARCH ARTICLES FROM SPANISH INTO ENGLISH: A CORPUS-BASED COMPARATIVE ANALYSIS OF THE GENRE

Translating research articles from any language into English is of paramount importance in the scientific community. However, before translating, it is necessary to ascertain the macrostructure of both source text and target text according to the genre conventions in each language. This article aims to prove whether research articles on the domain of Information and Technology Law published in Spanish share the Introduction-Method-Results-Discussion (IMRD) structure used in most articles written in English. More specifically, we focus on the section 'introduction' in order to study whether most articles have either the Create a Research Space (CARS) model (Swales, 1990) or the Open a Research Option (OARO) model (Swales, 2004). In previous studies with small corpora (Toledo Báez, 2009 and 2010), the results showed that the introductions CARS are much more frequent in English than in Spanish and the OARO structure is the most common in the Romance language. However, we need to prove this hypothesis with the intratextual comparative analysis of our bilingual, specialized, virtual, and representative (Corpas Pastor and Seghiri Domínguez, 2010/in press) comparable corpus consisting of a collection of 280 research articles on electronic commerce, 140 in Spanish and 140 in English. This article also pays attention to the possible macrostructural consequences when translating research articles from Spanish into English. This difference may have an impact on the translation of these texts because the translator may have to decide whether to keep the original features of the research article in Spanish or, on the contrary, to adapt the Romance text to the Anglo-Saxon conventions of research articles.

Torre Alonso, Roberto

Panel: 4. Lexicología y lexicografía basadas en córpora

THE PREFIX UN- IN THE FORMATION OF OLD ENGLISH NOUNS: COMBINATORIAL PROPERTIES AND CONSTRAINTS.

This paper aims at sheding light upon the morphological properties of the prefix un- in the formation of complex nouns in Old English. More concretely, this paper provides an exhaustive description of the combinatorial properties of the prefix as regards both the bases to which it may be attached and the affixes with which it can interact in recursive derivative processes. Thus this research is twofold On the one hand it explores the nature of the bases that admit derivation with un- as regards the lexical class to which they are adscribed. On the other hand, the morphological character of the bases is also discussed, whether simple or complex, thus allowing for the establishment of a set of affixes which can act in interaction with the prefix un-. In this respect, this research is supported by the works by Siegel (1979), Fabb (1988), Aronoff and Furthop (2002), Hay and Plagg (2004) or Lieber (2004) or Martin Arista (2010), which focus on the subject of affix combinations. Regarding the target language, this stage of the language is characterised by a rich inflectional system (Kastovsky 1992), and complex words clearly outnumber simple ones. Moreover, complex words are also used as bases for further derivational steps to operate, thus allowing for the existence of recursively derived words, which are a suitable field of study for the analysis of affix combinations. The analysed data have been retrieved from the lexical database Nerthus (www.nerthusproject.com), which contains over 30,000 predicates, and consists of a total of 162 predicates. Of these 153 present a nominal base, whereas 9 present a non-nominal base, which include 3 verbs and 6 adjectives. The reason for the existence of non-nominal bases is to be found in the analysis methodology and in the fragmentation of the surviving lexical stock of the period. Regarding the morphological complexity of the bases, only 35 are underived. Thus, nouns prefixed with un- present some degree of recursivity in 78.395% of the cases. Besides 71 of the predicates (43.827% of the total and 55.905% of the recursively derived nouns) present two affixes in the final word creation steps. According to the data, the prefix un- can combine with four different prefixes, namely ā-, for-, ful- and ge-, giving way to a total of 26 predicates, and with seven suffixes, those being -dōm, -en, -ere, -ing/-ung, -nes, -scipe, and –t, combining for a total of 45 nouns. The final part of the analysis tries to set the data against the Monosuffix Constraint, proposed by Aronoff and Fuhrhop (2002). These authors identify closing suffixes that do not allow for further derivation once they have been attached to a base. The data show the existence several that can occur once un- has occupied its place in the derivation The morphemes allowing for this combinatorial order are –dōm, -end, hād,-ing/-ung, -nes, and –t in some 93 predicates. These data don't allow claiming the status for un-, but show that the prefix is process final, as it admits no further prefixation.

Valverde-Mateos, Ana

Panel: 8. Los córpora y la adquisición y enseñanza del lenguaje

USO DE CORPUS ORALES DE APRENDIENTES PARA LA ENSEÑANZA DEL FRANCÉS COMO LENGUA EXTRANJERA

Además de las dificultades lógicas asociables a los distintos niveles de dominio de una L2, lo que suele ocurrir es que la lengua que se enseña a estos estudiantes no pertenece al mismo registro que la que utilizan los hablantes nativos. Cierto es que, cada día con mayor frecuencia, en las aulas, los estudiantes están expuestos a muestras variadas de expresión características de los distintos registros (desde los más coloquiales a los más formales). No es menos cierto que el habla espontánea de los nativos pertenece al registro coloquial, mientras que en los proceso de formación sistemática se hace especial hincapié en los modelos de lengua formal o considerada estándar, la cual se basa normalmente en la norma escrita, y resulta mucho más específica y artificial en relación con el francés oral espontáneo. Nuestra disposición es crear un estudio que pueda sensibilizar a los docentes de L2 a la necesidad de desarrollar estrategias de enseñanza basadas en el francés que utilizan los nativos. Estamos convencidos de que la observancia de materiales auténticos procedentes de entornos no nativos, podrá ofrecer a los docentes (generalmente, también no nativos), una manera diferente de abordar la lengua llena de posibilidades por explorar, y a los alumnos la oportunidad de disponer de un nuevo paradigma de estudio. Para ello, partimos de la hipótesis de que un corpus oral compuesto por entrevistas a hablantes de francés como segunda lengua, y organizado según los diferentes niveles del Marco Común de

Referencia Europeo para las Lenguas, puede ayudar a arrojar luz sobre aquellos aspectos más problemáticos de asimilar para estudiantes hispanohablantes, así como sobre sus errores más frecuentes. A nuestro entender, un análisis concienzudo del conjunto, permitirá ahondar en una nueva técnica de aprendizaje, que nos proponemos plasmar en un futuro en una plataforma de enseñanza online complementaria a las herramientas pedagógicas actuales. Retomamos así la idea de Braun (2005) que insiste en la necesidad de crear corpus para responder verdaderamente a las necesidades de profesores y aprendientes, enriquecerlas con ejercicios y combinar el estudio de textos completos con concordancias que muestren nuevos usos o ejemplos de uso y aspectos de la lengua menos evidentes y poco conocidos por los estudiantes, olvidando, en parte, los grandes y complejos corpus de referencia. En la presente comunicación, expondremos nuestra experiencia de partida, y detallaremos el proyecto que estamos llevando a cabo. De esta manera, se describirá el trabajo realizado para el diseño del corpus de hablantes de Francés Lengua Extranjera (FLE), destacando aspectos interesantes de su desarrollo e implementación. Así, nos detendremos en la concepción del corpus (generación del cuestionario para las diferentes entrevistas, tipos de hablantes, recogida de datos) y en procesos necesarios para su implementación (transcripción con la inclusión de comentarios y análisis de errores frecuentes) y su posterior análisis.

Varela Pérez, José Ramón

Panel: 6. Corpus y variación lingüística

NOT-NEGATION AND NO-NEGATION IN CONTEMPORARY SPOKEN BRITISH ENGLISH: A CORPUS-BASED STUDY

This contribution explores the interface between corpus linguistics, diachronic typology and usage-based approaches to the study of grammatical variation. I will address the alternation between two types of negation in contemporary spoken English involving non-specific indefinites under the scope of negation: NOT-negation (He did not see anything) and NO-negation (He saw nothing) (Tottie 1991a, 1991b, 1994). Historically, the possibility of variation between the older construction with NO-negation and the newer one with NOT-negation was only effective after the disappearance of ne and the rise of not as a marker of verbal negation at the end of the ME period (Jespersen 1917; Mazzon 2004). The demise of multiple negation of the type ne + verb + no/nothing/none, etc., brought about NO-negation: I ne saw nothing > I saw nothing. In addition, the new marker of verbal negation (not) could increasingly combine with negative polarity items such as any and the indefinite article (not…a/any/anything, etc.) (Shanklin 1988). There have not been many corpus-based studies of this topic. Most of them focus on contexts where variation is not possible and/or offer quantitative findings without further qualitative analysis of the data (e.g. Biber et al. 1999; Westin 2002; Peters 2008). Only Tottie (1991b) has offered a comprehensive study of this area although she relies on corpora dating back to the 1960s and the early 1970s. In this paper, I will use a sample of contemporary spoken British English taken from the British component of the International Corpus of English (ICE-GB), including conversations recorded in the early 1990s. In this regard, a comparison of my data with Tottie's (1991b) findings might reveal some evidence of on-going change in this area given the way changes from below in English grammar seem to have spread historically.

I will also address the impact of several internal factors on the choice between the two constructions, including some that have not yet been considered in the literature. Ultimately, the variation between NOT-negation and NO-negation must be placed against the backdrop of diachronic typology: the history of sentence negation in English (the so-called Jespersen's Cycle) and two competing typological tendencies that bear opposite results in the expression of negation: (a) the 'Neg First' principle, i.e. the universal psycholinguistic tendency for negative markers to be placed before the verb (Jespersen 1917; Horn 1989); and (b) the End-weight principle, i.e. the tendency to concentrate communicatively significant elements towards the second part of the sentence (Mazzon 2004).

Vea, Raquel

Panel: 4. Lexicología y lexicografía basadas en córpora

THE CORPUS PRODUCTIVITY OF OLD ENGLISH ADJECTIVAL COMPOUNDS WITH VERBAL BASE

This presentation aims at analyzing the productivity of Old English deverbal adjectival compounds. In this research, the productivity of a word-formation process in a historical language is based on an assessment of formal transparency and textual frequency, as put foward by Kastovsky (1992) and Lass (1994). The corpus of analysis has been retrieved from the lexical database of Old English Nerthus (www.nerthusproject.com), which turns out a total of 241 compounds, if spelling variants are disregarded. Focusing on the adjunct of the compound, three categories are involved: nominal adjunct (bordhæbbende 'shield-bearing': bord 'board'), adjectival adjunct (micelsprecende 'boasting': micel 1 'great, intense'), and adverbial adjunt (eftboren 'born again': eft 'again'). By type, the most frequent compounds are the following: æðelboren 'of noble birth' (29), hefigty:me 'heavy, grievous' (29), u:tancumen 1 'foreign, strange' (31), a:ncenned 'only-begotten' (74), frumcenned 'first-begotten' (77). Once the type anaysis has been carried out by means of the lexical database, the token analysis resorts to the Dictionary of Old English Web Corpus. The conclusions of the analysis go along the following lines. Regarding the sources, some difficulties arise in establishing the correspondences between lemmatized and unlemmatized forms and, as far as the question of token frequency is concerned, this type of compound is far more frequent in prose than in poetry.

Velasco Moreno, Mª Isabel

Panel: 8. Los córpora y la adquisición y enseñanza del lenguaje

INFLUENCIA DEL FEEDBACK EN EL ALUMNADO DE EDUCACIÓN PRIMARIA CON RESPECTO A SU PRODUCCIÓN ORAL EN LENGUA EXTRANJERA.

Numerosos investigadores desde Thorndike (1913) hasta nuestros días han reflexionado sobre el fenómeno del feedback desde distintas perspectivas (Butler y Winne, 1995; Klugger y DeNisi, 1996; Mackey, 2006; Hattie y Timperley, 2007; Brookhart, 2008; Huei-Hsin, 2009). La necesidad de encontrar características que permitan definir el feedback en clase ha sido una constante así como la búsqueda de una mayor eficacia del mismo aunque hay enormes diferencias entre ellos. Algunos dan una visión generalizada, a nivel teórico mientras que otros se centran en aspectos tan específicos como el feedback correctivo. Observamos, por otro lado, que la mayoría de estos trabajos han tenido lugar en contextos universitarios, habiendo escasez de estudios en niveles educativos inferiores. Nuestra investigación persigue detectar y analizar elementos reales que transmitan feedback en aulas de aprendices de lengua extranjera en un nivel de Educación Primaria para posteriormente investigar los efectos que hayan podido ocasionar en los discentes. En este estudio, tras la recopilación de un corpus procedente de clases regladas reales y actuales de grupos de alumnos andaluces de Educación Primaria, hemos aplicado un modelo de análisis del discurso comunicativo en el aula, basado en el modelo propuesto por la escuela de Birmingham (Sinclair y Coulthard, 1975); en el Análisis de Conversación (Tsui, 1992) y en la concepción tripartita del lenguaje (Poyatos, 1994). Hemos investigado cuestiones relativas a quién o quienes proporcionan FB; a quién/quienes se dirige el mismo; la selección de elementos verbales y no verbales utilizados para su realización en el movimiento de Follow up; interpretación que el alumnado hace del feedback recibido; distinción de tipos y efecto provocado. También se investiga los fragmentos de clase en los que aparece mayor índice de retroalimentación profundizando en el tipo de actividades realizadas cuando ésta tuvo lugar. Indudablemente, los resultados obtenidos mediante el análisis de este córpora van a ayudar a comprender cómo tiene lugar la adquisición de una segunda lengua para y este conocimiento permiten mejorar la enseñanza de L2.

Viberg, Åke

Panel: 5. Corpus, estudios contrastivos y traducción

IMPERSONAL CONSTRUCTIONS IN SWEDISH. A CORPUS-BASED CONTRASTIVE STUDY

Impersonal constructions have attracted a lot of attention recently from typologically oriented researchers (Siewierska 2008, Malchukov & Siewierska forthc.). This paper, which represents an extension of Viberg (2010), presents a corpus-based contrastive study of impersonals in Swedish based on the multilingual parallel corpus (MPC). At present, MPC consists of extracts from 22 novels in Swedish with translations of all texts into English, German, French, and Finnish. For some texts, translations also are included into Spanish, Italian, Dutch, Icelandic, Danish, and Norwegian. There is a total of around 600,000 words in the Swedish originals. In addition to this material, there are also some original texts in French and Finnish with translations into Swedish. Only part of the corpus has been analysed so far. (Author 2010 is based on five of the original texts in Swedish and their translations.) As a first step in the analysis, impersonals were identified with simple formal criteria. All occurrences of the Swedish generalized pronoun man and of non-referential det 'it' were extracted for further analysis. Swedish impersonals include a number of constructions with impersonal (dummy) det 'it' as subject: clefting, presentation, extraposition of finite and non-finite clauses and the impersonal passive. An analysis was also made of the distribution of impersonal verbs (and other predicates) across semantic fields. As a second step, this material was analysed from a functional point of view. It turned out that det appears as a formal subject (or placeholder) in agentless sentences or sentences with low agentivity, whereas man appears as an impersonal subject with general ('all of mankind') or vague reference. The individual constructions were also studied. From a contrastive perspective, it turns out that Finnish in many respects represents a different type than the other languages included in the study, but even if German, English, and French in many cases have rather direct structural equivalents to the Swedish impersonal constructions, the usage patterns differ in a striking way even between these languages. For example, in the material analyzed so far, there were 181 it-clefts in Swedish of the type It was Peter who came. It turned out that it-clefts and other clefts (pseudoclefts) were equally frequent as translations in English, but together these structures accounted for no more than 30% of the translations. For German, the proportion was even lower (20%). The highest correspondence was found in French with 43%, which is still rather low. Finnish does not have any direct structural corresponden to it-clefts and used other translations (including neutral sentences lacking any functional equivalent).

Voutilainen, Atro, Krister Linden and Tanja Purtonen

Panel: 1. Diseño, compilación y tipos de córpora

DESIGNING A DEPENDENCY REPRESENTATION AND GRAMMAR DEFINITION CORPUS FOR FINNISH

We outline the design and creation of a syntactically and morphologically annotated corpora of Finnish for use by the research community. We motivate a definitional, systematic "grammar definition corpus" as a first step in an three-year annotation effort to help create higher-quality, better-documented extensive parsebanks at a later stage. The syntactic representation, consisting of a dependency structure and a basic set of dependency functions, is outlined with examples. Reference is made to double-blind annotation experiments to measure the applicability of the newgrammar definition corpus methodology.

Westall, Debra

Panel: 2. Discurso, análisis literario y corpus

EL PAÍS NEWS REPORTS ON CHILDHOOD OBESITY: A TWELVE-MONTH CORPUS STUDY

Given the current global obesity epidemic and the media's coverage of this phenomenon over the past decade, researchers have begun to examine news reports on obesity through qualitative, quantitative, thematic, content and discourse analyses. Following the work of Lawrence (2004), Kim & Willis (2007) and Boero (2007) in the US, obesity news studies have been conducted in Australia (Udell & Mehta, 2008), Canada (Roy et al., 2007), Germany (Hilbert & Ried, 2009), Norway (Malterud & Ulriksen, 2009), Sweden (Sandberg, 2007) and the UK (Gough, 2007). Yet to my knowledge, there is no comparable study available for Spain. Therefore, the purpose of my ongoing research is to examine Spanish written press coverage of obesity, especially in regard to children (Author, in press; Author, 2010). This research is based on a specific corpus of 231 news items published between 01/01/2008 and 31/12/2008. This year was selected for study because in April of 2008 the national press ran headlines which blamed obesity for a child's death in Murcia, Spain. Using various combinations and synonyms of the key search term, obesidad infantil, all pertinent news items were extracted from the online archives of ABC, El Mundo and El País, the three leading national newspapers in Spain. After analyzing each item manually, only those results containing at least one direct reference to childhood/adolescent overweight/obesity were included in the final 231-item corpus: ABC (n=88; 38.1%), El Mundo (n=78; 33.8%), and El País (n=65; 28.1%) (total word count, approx. 135,932; 588 words/text). The present study will focus on the 65 items published in El País, the top-circulation daily in Spain. The El País articles tended to be longer (673 words/item) than those published in ABC and El Mundo (475 and 645 words/item, respectively). With an average of 5.4 items/month, the El País sample contains 9 (13.8%) opinion articles, 23 (35.4%) interpretative pieces, and 33 (50.8%) informative texts; the names of staff reporters or journalists appear on 51 (78.5%) of the items. The content analysis confirmed two types of thematic coverage: social and scientific. The social perspective frames news on public and private schemes to control or prevent obesity (16; 24.6%) as well as the implications of obesity in the lives of celebrities (8; 12.3%). The scientific frame is clear in news about obesity prevalence (6; 9.2%) or the causes (12; 18.5%) and health risks associated with childhood obesity (13; 20%). In brief, some 42 million children under five are overweight today and, according to the World Health Organization (2010), the majority will be overweight as adults; many will be diagnosed with diabetes or cardiovascular disease and, some, like the Spanish child in 2008, will die prematurely. The results of this research highlight the newsworthiness of the current childhood obesity epidemic in the leading Spanish daily and will provide relevant data for future studies of news framing and contemporary obesogenic discourse.

Wissik, Tanja

Panel: 1. Diseño, compilación y tipos de córpora

COMPILING SPECIALIZED CORPORA ACROSS LANGUAGE VARIETIES AND WORKING WITH THEM

The building and the analysis of specialized multilingual corpora on one hand and the building and the analysis of corpora of national varieties on the other hand are well established methods in translation studies and linguistics. But the analysis of specialized comparable corpora for national varieties is still in his infancy since most studies analyzing national varieties, especially for German, focus on general language and not on language for special purpose. In this paper the design and development of the so called UNI-Corpus will be described. The relevant corpora are compiled with a special regard to the institutional language used in the university systems in Austria, Germany and Switzerland. This paper will present the experience of developing these three comparable corpora and will discuss issues which arose when setting up the corpus, like the selection of texts, the size of the sub-corpora, regional distribution etc. Furthermore, the paper will discuss a case study to illustrate the application and the use of the UNI-Corpus, which can be used for comparative and contrastive studies, but the results of the analysis can also be used in translation studies and in the actual translation process.

You, Zixi

Panel: 3. Estudios gramaticales basados en córpora

A CORPUS-BASED EXAMINATION OF PERFECTIVE AUXILIARY SELECTION IN OLD JAPANESE

Auxiliary selection has received a great deal of attention among linguists who work on European languages, for example, Italian, French, Old Spanish, and so on, whereas very few studies have taken auxiliary selection in Asian languages into consideration. Washio (2002; 2004) argued that the perfect auxiliaries in Old Japanese (OJ) displayed a close distributional correspondence to the European auxiliaries H (HAVE) and B (BE); however, the full picture of the distribution and the underlying criteria for the auxiliary selection in OJ itself remained unclear and debatable. This paper, illustrating how a large corpus with a big amount of textural and grammatical data merits descriptive and analytical linguistic research, examines auxiliary selection in OJ by means of a newly completed OJ Corpus, a part of the VSARPJ Corpus that features a large amount of grammatical information encoded in pre-modern Japanese texts. The perfective auxiliary in OJ has two variants, '-(i)n-' and '-(i)te-'. As has been pointed out by Frellesvig (2010: 67), they belong closely together for the reasons that they are mutually exclusive, occupy the same position in a verb system, do not co-occur with the stative or the negative, and exhibit mostly the same inflected forms. The OJ Corpus consists of nearly all attested OJ texts, romanized and xml tagged with a wide range of linguistic information, e.g. orthography, part-of-speech, morphology, syntactic constituency, etc., following TEI conventions. (More recently, information about semantic role is also being added.) As part of the construction of the Corpus, I marked up, both automatically and manually, all the occurrences of perfective auxiliaries in OJ, and assigned ID numbers to all verbs preceding the perfectives (in the same word) according to the Lexicon of the Corpus. Xaira was used to extract the data from the Corpus. A comprehensive and exhaustive investigation was carried out on all verbs that precede perfective auxiliaries in both single and compound forms. In total, I found 199 verbs that only co-occurred with '-(i)n-', 112 verbs with '-(i)te-', and more interestingly, 18 verbs that could co-occur with both. Based on these lists, I looked at each verb in other contexts in the Corpus to investigate their syntactic behaviors, and also analyzed the interaction between semantic factors, namely, agentivity, volitionality, affectedness, and telicity. Results showed that agentivity and telicty played the most important roles in the auxiliary selection in OJ; furthermore, a strong tendency that transitive verbs and unergative verbs pattern with '-(i)te-' and unaccusative verbs pattern with '-(i)n-' was also observed. After a more closed examination of the verbs that selected both '-(i)n-' and '-(i)te-', compositional factors turned out to be the key that resulted in the syntactic variations in auxiliary selection, or, from another perspective, the extension of the domain of the selection.

Based on the largest and newest corpus for OJ language, this research contributes to a detailed description of verbs and auxiliary selection in OJ, benefiting future comparative studies of Eastern and Western languages, while also having implications for linguistic theory in general.