Agència de Gestió d'Ajuts Universitaris i de Recerca

MARIE CURIE ACTIONS

ATENCIÓ:

Aquest document només és vàlid per a annexar-lo, en format PDF, al formulari de sol·licitud BP 2010.

Les dades d'aquest annex podran ser modificades o ampliades per part de la persona sol·licitant fins a la data límit de presentació de les esmenes a les dades bàsiques de la sol·licitud, d'acord amb el que determinen les bases de la convocatòria. Transcorregut aquest termini, les sol·licituds seran avaluades amb la informació que hi consti.

PLEASE NOTE:

This document is only valid for attaching, in PDF format, to the BP 2010 application form.

The data of this annex may be modified or extended by the applicant until the deadline for submitting changes to the basic data of the application, in accordance with the requirements of the rules of the call for applications. Once this term has elapsed, the applications will be assessed with the information they contain.

Annex de sol·licitud de beques i ajuts Beatriu de Pinós
Annex to application form Beatriu de Pinós fellowships
BP 2010

Modalitat A: Beques per a estades de recerca postdoctorals fora de l'Estat espanyol
Modality A: Scholarships for postdoctoral research stays outside Spain

Dades de la persona candidata / *Details of the candidate*

Nom / *First name*	Primer cognom / *First surname*	Segon cognom / *Second surname*
Gemma	Boleda	Torrent
Tipus identificador / *Type of identification*	Número identificador / *Identification No.*	
	[redacted]	

Telèfon / *Tel. No.*	Mòbil / *Cell phone*	Correu electrònic / *E-mail address*
936612861	[redacted]	[redacted]

Generalitat de Catalunya
Departament d'Innovació, Universitats i Empresa
Comissionat per a Universitats i Recerca

1. Currículum de la persona candidata (màxim 8 fulls) / *Candidate's CV (maximum 8 pages)*

1.1 Formació acadèmica i experiència professional / *Higher education and professional experience*
1.2 Contractes d'R+D i participació en projectes finançats / *R&D contracts and participation in funded projects*
1.3 Publicacions i resultats científics / *Publications and scientific results*
1.4 Beques i altres tipus d'ajuts rebuts / *Grants and others fellowships received*
1.5 Estades a l'estranger / *Research stays abroad*
1.6 Participació en congressos i conferències / *Attendances to congresses and conferences*
1.7 Altres mèrits acadèmics i/o professionals / *Other academic and/or professional credits/recognitions*

1.1 *Higher education and professional experience*

1.1.1 Higher education

Universitat Pompeu Fabra, Spain
PhD in Cognitive Science and Language, 2007
DEA (equivalent to a M.A. degree) in Cognitive Science and Language, 2003
with Honours

Universitat Autònoma de Barcelona, Spain
B. A. (4-year degree), Spanish Philology, 2000
with Honours

Universität zu Köln, Germany
Studies in Natural Language Processing, 1997-98

1.1.1 Professional experience

- 2008-present: Post-doc researcher, Universitat Politècnica de Catalunya, Spain
- 2005-2007: Researcher, Barcelona Media Centre d'Innovació, Spain
- 2001-2007: Researcher, Universitat Pompeu Fabra, Spain
- 2000: Student assistant, Universitat Pompeu Fabra, Spain
- 2000: Student assistant, Artificial Intelligence Research Institute (IIIA, CSIC), Spain
- 1997-1998: Student assistant, Universität zu Köln, Germany

1.2 *R&D contracts and participation in funded projects*

2011-2013
OntoSem2: Natural language ontology and the semantic representation of abstract objects 2 (FFI2010-15006)
Funded by the Spanish government, €104,000
PI: Louise McNally, Universitat Pompeu Fabra

2011-2012
REDISIM: Modelado distribucional de las propiedades recursivas del significado (FFI2010-09464-E)
Funded by the Spanish government (MICINN), Acción Complementaria, subprograma EXPLORA, 8000€ (renegotiation underway)
PI: Louise McNally, Universitat Pompeu Fabra

2010-2012
KNOW2: Language understanding technologies for multilingual domain-oriented information access (TIN2009-14715-C04-04)
Funded by the Spanish government, €160,600.
PI: Jordi Turmo, Universitat Politècnica de Catalunya

2006-2009
KNOW: Developing large-scale multilingual technologies for language understanding (TIN2006-15049-C03-03)
Funded by the Spanish government, €100,000
PI: L. Padró, Universitat Politècnica de Catalunya

2007-2010
OntoSem: Natural language ontology and the semantic representation of abstract objects (HUM2007-60599/FILO)
Funded by the Spanish government, €109,100
PI: Louise McNally, Universitat Pompeu Fabra

2005-2006
Natural language ontology for reference to facts and eventualities (HA2005-0100 and HF2005-0177)
Trilateral *Acción Integrada* (Integrated Action)
Funded by the Spanish government, €10,831 and €11,080
PI: Louise McNally, Universitat Pompeu Fabra

2004-2007
METIS-II: Statistical Machine Translation using Monolingual Corpora: from Concept to Implementation (IST – FP6-003768)
Funded by the European Union, €1,000,000
General Coordinator: ILSP (Athens, Greece); PI at UPF: Toni Badia

2005-2008
ARQUITEXT: Arquitectura integrada para el tratamiento avanzado de textos (HUM2004-05321-C02-02)
Funded by the Spanish government, €25,920
PI: Toni Badia, Universitat Pompeu Fabra

2005-2008
NOCANDO: Construcciones no canónicas en el discurso oral: estudio transversal y comparativo (HUM2004-04463)
Funded by the Spanish government, €21.800
PI: Enric Vallduví, Universitat Pompeu Fabra

2001-2004
PrADo: Sistema de preparación automatizada de documentos (TIC2000-1681-C02-01)
Funded by the Spanish government, €36.000
PI: Toni Badia, Universitat Pompeu Fabra

1999-2001
Integración de técnicas y herramientas de tagging, parsing y unificación (PB98-1151)
Funded by the Spanish government
PI: Gabriel Amores, Universidad de Sevilla

1.3 *Publications and scientific results*

Note: most relevant publications boldfaced.

Dissertation

Boleda, G.. 2007. Automatic acquisition of semantic classes for adjectives. Ph.D. thesis, Universitat Pompeu Fabra. (Advisors: Toni Badia and Sabine Schulte im Walde.)

Books

Artstein, Ron, G. Boleda, Frank Keller, Sabine Schulte im Walde (eds). 2008. *Proceedings of the COLING Workshop on Human Judgements in*

***Computational Linguistics* (COLING). Manchester, UK: Coling 2008 Organizing Committee. ISBN 978-1-905593-49-1.**

Journal articles

Boleda, G., M. Cuadros, C. España-Bonet, Maite Melero, L. Padró, M. Quixal, Carlos Rodríguez. 2009. Primera Jornada del Procesamiento Computacional del Catalán. *Revista de Procesamiento del Lenguaje Natural* 43: 387-388. ISSN: 1135-5948.

Boleda, G., M. Cuadros, C. España-Bonet, L. Padró, Maite Melero, M. Quixal, Carlos Rodríguez. 2009. El català i les tecnologies de la llengua. *Llengua, Societat i Comunicació* 7: 20-26. ISSN: 1697 5928.

Boleda, G., M. Cuadros, C. España-Bonet, Maite Melero, L. Padró, M. Quixal, C. Rodríguez. 2009. Sobre la I Jornada del Processament Computacional del Català. *Llengua i Ús* 45: 23-32. ISSN: 1134-7724.

Boleda, G., S. Schulte im Walde, T. Badia. 2008. An Analysis of Human Judgements on Semantic Classification of Catalan Adjectives. *Research on Language and Computation* 6(3): 247-271.

Boleda, G. 2008. Emulant els infants: induint propietats lingüístiques a partir de dades empíriques. *Revista de Catalunya*, 235, pp. 33-40. ISSN 0213-5876.

Badia, T., G. Boleda, M. Melero, A. Oliver. 2005. El proyecto METIS-II. *Revista de Procesamiento del Lenguaje Natural*. ISSN 1135-5948, 35, pp. 443-444.

Oliver, A., T. Badia, G. Boleda, M. Melero. 2005. Traducción automática estadística basada en n-gramas. *Revista de Procesamiento del Lenguaje Natural*. ISSN 1135-5948, 35, pp. 77-84.

Mayol, L., G. Boleda, T. Badia. 2005. Automatic acquisition of syntactic verb classes with basic resources. *Language Resources and Evaluation*, 39(4):295-312

Alsina, Àlex, T. Badia, G. Boleda, Stefan Bott, Àngel Gil, M. Quixal, Oriol Valentín. 2002. CATCG: un sistema de análisis morfosintáctico para el catalán. *Revista de Procesamiento del Lenguaje Natural*, 29, Sept. 2002, pp. 309-310. ISSN: 1135-5948

Badia, T., G. Boleda, Jenny Brumme, C. Colominas, Mireia Garmendia, M. Quixal. 2002. BancTrad: un banco de corpus anotados con interficie web. *Revista de Procesamiento del Lenguaje Natural*, 29, septiembre 2002, pp. 293-294. ISSN: 1135-5948

Badia, T., G. Boleda, M. Quixal. 2001. Curso sobre Tecnologías de la lengua (segunda edición). *QUARK, Ciencia, Medicina, Comunicación y Cultura*, 21, Jul - Sept. 2001. pp. 14-16. ISSN: 1135-8521

Book chapters

McNally, Louise and G. Boleda. 2004. Relational adjectives as properties of kinds. In Olivier Bonami and Patricia Cabredo Hofherr (eds.) *Empirical Issues in Syntax and Semantics 5*, pp. 179-196

Articles in refereed conference proceedings

Arsenijevic, B., B. Gehrke, G. Boleda, L. McNally. In press. Ethnic adjectives are proper adjectives. In *Proc. of 46th Annual Meeting of the Chicago Linguistic Society*, Chicago, IL, USA.

Sánchez Marco, C., G. Boleda, J. M. Fontana. In press. Propuesta de codificación de la información paleográfica y lingüística para textos diacrónicos del español. Uso del estándar TEI. In *Actas del Contreso Internacional Tradición e innovación: Nuevas perspectivas para la edición y el estudio de documentos antiguos*, Madrid, 11-13 November 2009.

Melero, M., G. Boleda, M. Cuadros, C. España-Bonet, L. Padró, M. Quixal, C. Rodríguez, R. Saurí. 2010. Language technology challenges of a 'small' language (Catalan). In *Proc. of LREC 2010*, Valletta, Malta. ISBN 2-9517408-6-7.

Peris, A., M. Taulé, G. Boleda, H. Rodríguez. 2010. ADN-classifier: Automatically assigning denotation types to nominalizations. In *Proc. of LREC 2010*, Valletta, Malta. ISBN 2-9517408-6-7.

Reese, S., G. Boleda, M. Cuadros, L. Padró, G. Rigau. 2010. Wikicorpus: A word-sense disambiguated multilingual Wikipedia corpus. In *Proc. of LREC 2010*, Valletta, Malta. ISBN 2-9517408-6-7.

Sánchez-Marco, C., G. Boleda, J.M. Fontana, J. Domingo. Annotation and representation of a diachronic corpus of Spanish. In *Proc. of LREC 2010*, Valletta, Malta. ISBN 2-9517408-6-7.

Sanromà, R. and G. Boleda. 2010. The Database of Catalan Adjectives. In *Proc. of LREC 2010*, Valletta, Malta. ISBN 2-9517408-6-7.

Boleda, G., S. Schulte im Walde, T. Badia. 2007. Modelling Polysemy in Adjective Classes by Multi-Label Classification. In *Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 171-180.

Boleda, G., S. Bott, C. Castillo, R. Meza, T. Badia, V. López. 2006. CUCWeb: a Catalan corpus built from the Web. In *Proc. of the Second Workshop on the Web as a Corpus at EACL (EACL)*. Trento, Italy, April 2006.

Badia, T., G. Boleda, M. Melero, A. Oliver An n-gram approach to exploiting a monolingual corpus for Machine Translation. In *Proc. of the Second Workshop on Example-based Machine Translation*, MT Summit X, Phuket, Thailand, 16 September 2005.

Boleda, G., T. Badia, S. Schulte im Walde. 2005. Morphology vs. Syntax in Adjective Class Acquisition. In *Proc. of the ACL-SIGLEX 2005 Workshop on Deep Lexical Acquisition (ACL)*, June 30, Ann Arbor, USA.

Mayol, L., G. Boleda, T. Badia. 2005. Automatic learning of syntactic verb classes. In *Proc. of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, pp. 92-97, Feb 28th-March 1st, Saarbrücken, Germany.

Boleda, G., T. Badia and E. Batlle. 2004. Acquisition of Semantic Classes for Adjectives from Distributional Evidence. In *Proc. of the 20th International Conference on Computational Linguistics (COLING)*, pp. 1119-1125, Geneva, Switzerland. ISBN:1-932 432-48-5

Padó, S. and G. Boleda. 2004. The Influence of Argument Structure on Semantic Role Assignment. In *Proc. of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, July 25-26, Barcelona, Spain. ISBN: 1-932432-36-1

Boleda, G. and L. Alonso. 2003. Clustering Adjectives for Class Acquisition. In *Proc. of the 10th Conference of The European Chapter of the Association for Computational Linguistics Student Research Workshop (EACL)*, pages 9-16, Budapest, Hungary. ISBN: 1-932432-01-9

Alsina, À., T. Badia, G. Boleda, S. Bott, À. Gil, M. Quixal, O. Valentín. 2002. CATCG: a general purpose parsing tool applied. In *Proc. of Third International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Spain, Vol. III, pp. 1130-1135. ISBN: 2-9517408-0-8

Badia, T., G. Boleda, C. Colominas, M. Garmendia, A. González, M. Quixal. 2002. BancTrad: a web interface for integrated access to parallel annotated corpora. In *Proc. of the First International Workshop On Language Resources For Translation Work And Research* held during the 3rd LREC Conference (LREC 2002), Las Palmas, Spain, 28 May 2002.

Badia, T., G. Boleda, M. Quixal, E. Bofias. 2001. A modular architecture for the processing of free text. In *Proc. of the Workshop on Modular Programming applied to Natural Language Processing at EUROLAN 2001*, Iasi, Romania, August 2001. pp. 11-18

Manuscripts

G. Boleda, S. Schulte im Walde, T. Badia. In prep. Modeling regular polysemy: A study in the semantic classification of Catalan adjectives. Under review at *Computational Linguistics*.

Corral, A., R. Ferrer i Cancho, G. Boleda, A. Diaz-Guilera. In prep. Universal Complex Structures in Written Language. Preliminary version available at http://arxiv.org/abs/0901.2924.

1.4 Awards, grants and fellowships received

- Post-doc contract of Spanish government, *Juan de la Cierva* programme, 2008-2011.
- PhD grant of *Fundación Caja Madrid*, 2005-2006.
- PhD grant of *Generalitat de Catalunya* (Catalan Government), 2001-2004.
- Extraordinary Spanish Philology Degree Award by the Universitat Autònoma de Barcelona, 2000.
- Honorable Mention of the National Bachelor Degree Awards by the Spanish Government, 2001.
- Grant for the Introduction to Research of the *Consejo Superior de Investigaciones Científicas*. Artificial Intelligence Research Institute (IIIA, CSIC), September-December 2000.
- Sócrates-Erasmus studentship, Universität zu Köln, Cologne, Germany, 1997-1998.
- Studentship of the DAAD (German Academic Exchange Service) for a language course at U. Gesamthochschule Essen, Germany, 1997.
- Studentship of the Spanish Education Ministry for the first B.A. academic year because of Honours in secondary school.

1.5 *Research stays abroad*

Center: *Sprachliche Informationsverarbeitung* department, University of Cologne
Location: Cologne, Germany
Dates: September 1997 - May 1998
Topic: Machine Translation
Funding: EU Erasmus programme.

Center: Institute for Phonetics and Computational Linguistics (CoLi), Saarland University
Location: Saarbrücken, Germany
Dates: May-July 2003
Topic: Computational lexical semantics
Funding: Generalitat de Catalunya and SALSA project.

Center: Institute for Phonetics and Computational Linguistics (CoLi), Saarland University
Location: Saarbrücken, Germany
Dates: November-December 2004
Topic: Computational lexical semantics
Funding: Generalitat de Catalunya.

Center: *Institut für Maschinelle Sprachverarbeitung*, Stuttgart University
Location: Stuttgart, Germany
Dates: April-August 2010
Topic: Computational lexical semantics
Funding: PASCAL2 European Network of Excellence (€6,000) and SFB 732 project (€4200).

1.6 *Attendances to congresses and conferences*

Note: only references to conference participation (as opposed to attendance) are included; and the ones that gave rise to publications are already included in Section 1.3, so they are not repeated below.

Gehrke, B., B. Arsenijevic, G. Boleda, L. McNally. 2010. Ethnic adjectives are proper adjectives. *46th Annual Meeting of the Chicago Linguistic Society*, Chicago, IL, USA, 8-10 abril. *Talk.*

Arsenijevic, B., G. Boleda, B. Gehrke, L. McNally. 2010. Unifying the semantics for "thematic" and "classificatory" uses of ethnic adjectives. *8èmes Journées Sémantique et Modélisation*, LORIA-INRIA, Nancy, France, 25-26 marzo. *Talk.*

Boleda, G., Á. Corral, R. Ferrer i Cancho, A. Díaz-Guilera. 2009. From word recurrence patterns to cognitive mechanisms. *15th Annual Conference on Architectures and Mechanisms for Language Processing (AMLAP).* Barcelona, 7-9 septiembre. *Poster.*

Berndt, D., G. Boleda, B. Gehrke, L. McNally. 2009. Nominalizations and nationality expressions: A corpus analysis. *Corpus Linguistics 2009.* Liverpool, UK, 20-23 julio. *Talk.*

Boleda, G., Á. Corral, R. Ferrer i Cancho, A. Díaz-Guilera. 2009. Word distance distribution in literary texts. *Corpus Linguistics 2009.* Liverpool, UK, 20-23 julio. *Talk.*

Boleda, G. 2009. Uso de PLN en otras disciplinas . *III Jornadas PLN-TIMM: Modelos y técnicas para el acceso a la información multilingüe y multimodal en la web .* 5-6 febrer, Colmenarejo, Madrid. *Talk.*

Corral, Á., R. Ferrer i Cancho, G. Boleda, A. Díaz-Guilera. 2008. Universality classes and community structure in word recurrence. *BCNet Workshop - trends and perspectives in complex networks*. 10-12 diciembre, Barcelona. *Talk.*

Boleda, G., S. Bott, R. Meza, C. Castillo, T. Badia, V. López. 2005. Usant la web per estudiar el català. *III Jornades sobre el català a les noves tecnologies*, 14-16 abril, Barcelona. *Talk.*

Mayol, L., G. Boleda, T. Badia. 2005. Automatic learning of syntactic verb classes. *Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, 28 feb-1 marzo, Saarbrücken, Alemania. *Talk.*

Boleda, G., S. Bott, B. Poblete, C. Castillo, M.E. Fuenmayor, T. Badia, V. López. 2004. CuCWeb: un corpus del català construït a partir de la web. *II Congrés Online de l'Observatori per a la Cibersocietat*, Barcelona. *Online presentation.*

Colominas, Carme y Gemma Boleda. 2004. The extraction of translationally relevant information from small ad-hoc corpora. *Third International Conference on Corpus Use and Learning to Translate (CULT-BCN)*, Barcelona. *Talk.*

McNally, Louise y Gemma Boleda. 2003. Relational Adjectives as Properties of Kinds. *CSSP 2003 (The Fifth Syntax and Semantics Conference in Paris)*, 2-4 Octubre, París, Francia. *Talk.*

Padó, Sebastian y Gemma Boleda. 2003. Towards Better Understanding of Automatic Semantic Role Assignment. *Prospects and Advances in the Syntax/Semantics Interface*, Nancy, Francia. *Talk.*

Alonso, Laura and Gemma Boleda. 2002. An Approach to Catalan Adjective Lexical Classes by Clustering. Workshop on Quantitative Investigations for Theoretical Linguistics, Osnabrück, Alemania, 3-5 Octubre 2002. Talk.

1.7 *Other academic and/or professional credits/recognitions*

1.7.1 Student supervision

- Co-supervisor (with Josep Maria Fontana), Cristina Sánchez-Marco: El desarrollo del perfecto en español: un estudio de corpus, 2010, PhD thesis project, Universitat Pompeu Fabra.
- Supervisor, Samuel Reese: *WikiNet: Construction d'une ressource lexico-sémantique multilingue à partir de Wikipedia*, Master's thesis, 2009, ISAE (Institut Supérieur de l'Aéronautique et de l'Espace).
- Supervisor, Daniel Berndt, Student assistant at Universitat Politècnica de Catalunya during B. A. studies at Universität Osnabrück, Germany, October 2008 – February 2009.

1.7.2 Professional services and other activities

A. Organization of scientific activities

- Co-organiser, *Jornada del Processament Computacional del Català*, Barcelona, March 2009.
- Co-organiser, *Nanoworkshop on statistical physics and linguistics*, Barcelona, March 2009.
- Co-organiser, *Coling 2008 workshop on human judgements in Computational Linguistics*, Manchester, UK, July 2008.
- Organiser, reading group on Computational Semantics at Universitat Politècnica de Catalunya, 2008-present.
- Organiser, seminar GLiCom's seminar on Computational Linguistics, several years.

B. Reviewing

- **Journal reviewing:** Language Resources and Evaluation (2008), Corpora (2008).
- **Conference reviewing:** ACL-HLT 2011 (Portland, Oregon, USA) COLING 2010 (Beijing, China), LREC 2010 (Valletta, Malta), EMNLP 2009 (Singapore), SEPLN 2009 (Donostia, Spain), NODALIDA 2009 (Odense, Denmark), EACL 2009 (Athens, Greece), ACL 2008 (Columbus, Ohio, USA).
- **Workshop reviewing:** First Workshop on Computational Neurolinguistics at NAACL-HLT (Los Angeles, USA), Compositionality and Distributional Semantic Models (Workshop organized as part of ESSLLI 2010, Copenhagen, Denmark), CBA 2010 (Corpus-Based Approaches to Paraphrasing and Nominalization 2010; Barcelona, Spain), ESSLLI 2008 Distributional Lexical Semantics Workshop (Hamburg, Germany), Student Research Workshop at ACL 2007 (Prague, Czech Republic), Workshop on Contextual Information in Semantic Space Models (CoSMo 2007): Beyond Words and Documents (Roskilde University, Denmark), Student Research Workshop at EACL 2006 (Trento, Italy), Penn Linguistics Colloquium (2005, 2006).

1.7.3 Invited talks

- 2/06/2010: Word Sense Disambiguation and regular polysemy. *Institutsversammlung*, Institut für Maschinelle Sprachverarbeitung, Stuttgart, Germany.
- 19/03/2010: Computational Feedback to Linguistics: A study in the semantic classification of Catalan adjectives. *Nancy NLP Seminar*, INRIA-Lorraine, France.
- 14/11/2007: Automatic acquisition of semantic classes for adjectives. *Natural Language Processing Seminar*, Universitat Politècnica de Catalunya, Barcelona, Spain.
- 11/11/2004: Acquisition of Semantic Classes for Adjectives. *Colloquium of the International Post-Graduate College on Language Technology and Cognitive Systems*, Universität des Saarlandes, Saarbrücken, Germany.

- 25/11/2004: A Quantitative Approach to the Lexical Semantics of Adjectives. *Computational Linguistics Colloquium*, Universität des Saarlandes, Saarbrücken, Germany.
- 10/12/2004: Acquiring Semantics Classes for Adjectives through Clustering. *Computational Linguistics Seminar*, King's College, London, Great Britain.
- 28/06/2005: Adquisició de classes semàntiques adjectivals. *III Workshop of the PhD Program in Cognitive Science and Language: "Acquisition"*, Barcelona, Spain.

1.7.4 Teaching experience

Courses at the Departament de Traducció i Ciències del Llenguatge (Department of Translation and Language Sciences), Universitat Pompeu Fabra:

- *Lingüística Computacional I* (Computational Linguistics I), 2001-02 and 2002-03;
- *Lingüística Computacional II* (Computational Linguistics II), 2001-02;
- *Sistemes de Traducció Automàtica* (Machine Translation Systems), 2001-02 and 2002-03;
- *Informàtica Aplicada a la Traducció* (Informatics Applied to Translation), 2003-04;
- *Introducció a la Lingüística Computacional: Aprendre a programar en Prolog* (Introduction to Computational Linguistics: Learning how to program in Prolog), 2006-07 and 2009-2010;
- *Noves Tecnologies i Traducció* (New Technologies and Translation), 2006-07;
- *Pragmàtica i Semàntica Computacionals* (Computational Pragmatics and Semantics), 2008-09.

Co-taught with Stefan Evert: course on Computational Lexical Semantics at *21st European Summer School in Logic, Language and Information* (ESSLLI 2009), Bordeaux, France, July 20-31.

---

2. Historial científic del grup d'acollida de fora de l'estat espanyol (màxim 5 fulls) / *Expertise in the field of the host outgoing institution (maximum 5 pages)*

2.1 Breu descripció del grup d'acollida / *Brief description of the hosting research group*
2.2 Principals publicacions i resultats científics obtinguts en els darrers cinc anys / *Track record of significant research achievements in the last five years*

2.1 Brief description of the hosting research group

The core areas of the present proposal are semantic theory and computational linguistics, and complementary fields are cognitive science, statistics, and artificial intelligence. The **Department of Linguistics at the University of Austin at Texas** (henceforth, UT) provides an ideal environment to carry it out. The Linguistics Department at UT comprises 8 full professors, 7 assistant professors, and 3 associate professors, as well as 57 graduate students and a number of researchers. The department has three high-profile professors: Ian Hancock, who represented the Romani people at the United Nations and served as a member of the U.S. Holocaust Memorial Council under President Bill Clinton; Nora England, who was a MacArthur Fellow (1993-1998) and is the founding director of the Center for Indigenous Languages of Latin America; and Hans Kamp, an eminent researcher in semantics and philosophy of language. It also includes, among others, researchers in syntax (Stephen Wechsler, John Beavers, Jason Baldridge), semantics (David Beaver), psycholinguistics (Colin Bannard), documentary linguistics (Nora England, Anthony Woodbury, Patience Epps), and computational linguistics (Jason Baldridge, Katrin Erk, Colin Bannard).

The Department fosters interdisciplinary collaboration with further linguists in language departments. Especially relevant are Nicholas Asher, from the Department of Philosophy, who is an expert in formal semantics, Hans Boas, from the Department of Germanic Studies, who has been working on Frame semantics, and Lars Hinrichs in the English department, who has been doing corpus-linguistic studies and statistical analyses using the R statistics package. Also relevant are further computational linguists in other departments, such as Ray Mooney in the Computer Science Department and Matt Lease in the School of Information. These and other researchers are agglutinated in the **Computational Linguistics lab**, which provides intense interaction between the departments, for instance through a common computational linguistics reading group that meets every two weeks.

Of special relevance for this proposal is also the recently founded **Division of Statistics and Scientific Computation**, with which K. Erk, J. Baldridge, R. Mooney and M. Lease are all associated. The division coordinates a suite of advanced courses designed to address the needs of different disciplines. It also provides consulting services for UT students, faculty and staff, bring prominent faculty to campus as part of the Distinguished Lecture Series, and yearly organize a Summer Statistics Institute.

2.2 Track record of significant research achievements in the last five years

About the responsible scientist: Katrin Erk will be G. Boleda's responsible scientist at UT. She is an assistant professor (tenure-track position) at UT since 2006. She received her PhD at Saarland University in 2002, under the supervision of Gert Smolka and Manfred Pinkal, on *Parallelism constraints in underspecified semantics*, and her specialization is in lexical semantics, computational semantics, corpora, statistical natural language processing, and machine learning. She has recently received a prestigious CAREER award of $433,449 from the National Science Foundation (USA) for a project entitled *Word meaning: beyond dictionary senses*.

Dr. Erk is an essential backbone of the present proposal, since she is leading a development in the computational lexical semantics field from a sense enumeration model to a graded, feature-based model for word meaning in context, and its computational modelling. Her recent keynote talk at the ACL Workshop on Geometrical Models of Natural Language Semantics is an example of this leading role. Dr. Erk's recent work includes research on feature-based computational models for word meaning in context as well as annotation

studies on word sense and polysemy. In collaboration with Dr. Mooney, she is studying the integration of deep sentence semantics with graded, feature-based models for word meaning. She has two ongoing international collaborations on topics central to the proposal, one with S. Padó (U. Heidelberg) and one with D. McCarthy (Lexical Computing Ltd., Brighton). A representative list of recent projects and publications by her is provided below, together with those of other UT faculty.

About the Computational Linguistics lab: Faculty at the Computational Linguistic lab carry out research, among other topics, on:

- models of word meaning (modeling vagueness and polysemy, with a focus on lexical composition);
- automatic semantic analysis (shallow semantic parsing –i.e. automatic predicate/argument structure analysis for free text–, spatial and temporal analysis);
- formal aspects of the semantics and pragmatics of natural languages;
- discourse analysis (including SDRT and Discourse Modes);
- Combinatory Categorial Grammar (CCG, an efficiently parseable, yet linguistically expressive grammar formalism);
- language acquisition and cognitive aspects of learning;
- computational aspects of all the above (inference mechanisms, machine learning methods, implementation and testing).

Below is a selection of projects and publications by these faculty in the period 2006-2010 (please note that only items that are related to the present proposal are listed).

1) Projects

2010-2013: National Science Foundation. Perceptually Grounded Learning of Instructional Language (IIS-1016312). PI: Ray Mooney ($450,000).

2010-2011: Longhorn Innovation Fund for Technology. Enabling Data-Intensive Research and Education at UT Austin via Cloud Computing. PI: Matthew Lease, iSchool; Co-PIs: Weijia Xu, Texas Advanced Computing Center, Jason Baldridge ($94,000).

2010-2012: New York Community Trust. Spatial and Temporal Analysis of Multilingual Texts, PI: Jason Baldridge, co-PIs: David Beaver, Katrin Erk ($120,000).

2010-2012: National Science Foundation. Semantics and Pragmatics of Projective Meaning across Languages, PI: David Beaver. ($102,000 at UT Austin, lead institution of $399,200 total collaborative grant with Carnegie Mellon University and The Ohio State University; award confirmed, but final contract still pending).

2009-2014: National Science Foundation CAREER program. CAREER: Word meaning: beyond dictionary senses (NSF IIS 0845925). PI: Katrin Erk ($433,449).

2009-2012: National Science Foundation. Modeling Discourse and Social Dynamics in Authoritarian Regimes, PI: David Beaver ($349,676 of $1,850,000 total collaborative grant with University of Memphis and Cornell University).

2009-2010: Google Grant Program. Unsupervised Induction of Semantic Lexicons Handling Both Synonymy and Polysemy. PI: Ray Mooney ($50,000).

2008-2011: Army Research Office, Multi-disciplinary University Research Initiative (through subcontract from the University of Washington). A Unified Approach to Abductive Inference (W911NF-08-1-0242). PI: Ray Mooney ($378,267).

2008-2010: New York Community Trust. Multilingual Interpretation of Temporal Expressions in Text, PI: Jason Baldridge, co-PIs: David Beaver, Katrin Erk ($120,000).

2007-2010: National Science Foundation. Learning Language Semantics from Perceptual Context (IIS-0712097). PI: Ray Mooney ($443,535).

2007-2008: National Science Foundation. Reducing annotation effort in the documentation of languages using machine learning and active learning (NSF BCS 065198), PI: Jason Baldridge, co-PI: Katrin Erk ($79,106).

2007: Google Grant Program. Global Extraction of Semantic Relations from Text Corpora by Learning from Weak Supervision. PI: Ray Mooney ($60,000).

2006-2009: National Science Foundation. Autonomic Systems: Integrating Machine Learning with Computer Systems (CNS-0615104). PI: Emmett Witchel, co-PIs: Ray Mooney, Peter Stone, Yin Zhang, Vitaly Shmatikov ($880,000).

2006-2008: Information and Intelligent Systems, National Science Foundation. Extracting and Using Discourse Structure to Resolve Anaphoric Dependencies: Combining Logico-Semantic and Statistical Approaches (NSF IIS 0535154). PI: Nicholas Asher, co-PI: Jason Baldridge ($249,869).

2005-2009: Defense Advanced Research Projects Agency (through subcontract from Institute for Study of Learning and Expertise). Transfer Learning in Integrated Cognitive Systems (FA8750-05-2-0283). PI: Ray Mooney, co-PI: Peter Stone ($953,254).

2004-2006: Defense Advanced Research Projects Agency (through subcontract from Lockheed Inc.). Architecture for Cognitive Information Processing. PI: Emmett Witchel, co-PIs: Ray Mooney, Peter Stone, Michael Dahlin, Risto Miikkulainen, Doug Burger, Steve Keckler ($700,000).

2) Publications

To appear/in press

Asher, N. (to appear). *Lexical Meaning in Context*. Cambridge: Cambridge University Press.

Baldridge, J. (to appear). Categorial Grammar. To appear in Patrick Colm Hogan (ed.), *Cambridge Encyclopedia of the Language Sciences*.

Beavers, J. (to appear). Lexical Aspect and Multiple Incremental Themes. In V. Demonte and L. McNally (eds.), *Telicity and Change of State in Natural Language: Implications for Event Structure*. Oxford: Oxford University Press.

Beavers, J. (to appear). An Aspectual Analysis of English Ditransitive Verbs of Caused Possession. *Journal of Semantics*.

Beavers, J. (in press). The structure of lexical meaning: Why semantics really matters. *Language.*

Erk, K., S. Pado, and U. Pado (to appear). A Flexible, Corpus-driven Model of Regular and Inverse Selectional Preferences. *Computational Linguistics* 36(4), December 2010.

Geurts, B. and D. Beaver (to appear). Presupposition. In Maienborn, C., K. von Heusinger, and P. Portner (eds.), *Semantics: An. International Handbook of Natural Language Meaning.* Berlin: De Gruyter.

Pado, S., and K. Erk (in press). Translation Shifts and Frame-Semantic Mismatches: A Corpus Analysis. *International Journal of Corpus Linguistics*.

2010

Beaver, D. (2010). Have you Noticed that your Belly Button Lint Colour is Related to the Colour of your Clothing? In Rainer Bäuerle, Uwe Reyle, and Thomas E. Zimmermann (eds.), *Presuppositions and Discourse: Essays offered to Hans Kamp*. Oxford: Elsevier, 65-99.

Beavers, J., B. Levin, and S.W. Tham (2010). The typology of motion expressions revisited. *Journal of Linguistics,* 46:331-377.

Boas, H. (ed.) (2010). *Contrastive Studies in Construction Grammar*. John Benjamins.

Calhoun, S., J. Carletta, J. Brenier, N. Mayo, D. Jurafsky, M. Steedman, and D. Beaver (2010). The NXT-format Switchboard Corpus: A Rich Resource for Investigating the Syntax, Semantics, Pragmatics and Prosody of Dialogue. *Language Resources and Evaluation*, 44:387-419.

Chen, D., Kim, J.H., and Mooney, R.J. (2010). Training a Multilingual Sportscaster: Using Perceptual Context to Learn Language. *Journal of Artificial Intelligence Research*, 37:397-435.

Erk, K., and S. Padó (2010). Exemplar-Based Models for Word Meaning in Context. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, Uppsala, Sweden.

Matthews, D. E. and C. Bannard (2010). Children's production of unfamiliar word sequences is predicted by positional variability and latent classes in a large sample of child-directed speech. *Cognitive Science*, 34(3):465-488.

Ravi, S., J. Baldridge, and K. Knight (2010). Minimized models and grammar-informed initialization for supertagging with highly ambiguous lexicons. In *Proceedings of 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, Uppsala, Sweden.

Reisinger, J. and Mooney, R.J. (2010). Multi-Prototype Vector-Space Models of Word Meaning. In *Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 109-117, Los Angeles, CA.

2009

Asher, N., J. Dever, C. Pappas (2009). Supervaluations Debugged. *Mind*, 118(472):901-933. Oxford University Press.

Bannard, C., E. Lieven and M. Tomasello (2009). Modeling Children's Early Grammatical Knowledge. *Proceedings of the National Academy of Sciences*, 106:17284-17289.

Boas, H.C. (2009). *Multilingual FrameNets in Computational Lexicography: Methods and Applications*. Berlin: Mouton de Gruyter.

Erk, K. (2009). Representing words as regions in vector space. In *Proc. of the Thirteenth Conference on Computational Natural Language Learning* (CoNLL), Boulder, CO.

Erk, K. and D. McCarthy (2009). Graded word sense assignment. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Singapore.

Erk, K., D. McCarthy, and N. Gaylord (2009). Investigations on Word Senses and Word Usages. In *Proceedings of the Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP)*, Singapore.

Ge, R. and Mooney, R.J. (2009). Learning a Compositional Semantic Parser using an Existing Syntactic Parser. In *Proceedings of the Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP)*, 611-619, Singapore.

Kulis, B., Basu, S., Dhillon, I., and Mooney, R.J. (2009). Semi-supervised Graph Clustering: A Kernel Approach. *Machine Learning*, 74(1):1-22.

2008

Baldridge, J. (2008). Weakly supervised supertagging with grammar-informed initialization. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*, 57-64, Manchester, UK.

Beaver, D. and B. Clark (2008). *Sense and Sensitivity: How Focus Determines Meaning*. Blackwell, Oxford.

Beavers, J. (2008). On the Nature of Goal Marking and Delimitation: Evidence from Japanese. *Journal of Linguistics*, 44:283-316.

Culo, O., K. Erk, S. Padó and S. Schulte im Walde (2008). Comparing and Combining Semantic Verb Classifications. *Language Resources and Evaluation*, 42(3):265-291.

Erk, K. and S. Padó (2008). A structured vector space model for word meaning in context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Waikiki.

Hoyt, F. and J. Baldridge (2008). A Logical Basis for the D Combinator and Normal Form Constraints in Combinatory Categorial Grammar. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*, 326-334, Columbus, OH.

Mooney, R.J. (2008). Learning to Connect Language and Perception. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI)*, Senior Member Paper, 1598-1601, Chicago, IL.

2007

Asher, N. (2007). A large view of linguistic content. *Pragmatics & Cognition*, 15(1):17-39. John Benjamins.

Asher, N. and E. McCready (2007). 'Might', 'Would', 'Could' and a Compositional Account of Counterfactuals. *Journal of Semantics*, 24:1-37. Oxford: Oxford University Press.

Beaver, D., B. Clark, E. Flemming, T. F. Jaeger, and M. Wolters (2007). When Semantics Meets Phonetics: Acoustical Studies of Second Occurrence Focus. *Language*, 83(2):245-276.

Denis, P. and J. Baldridge (2007). Joint determination of anaphoricity and coreference resolution using integer programming. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, 236-243. Rochester, NY.

Erk, K. (2007). A Simple, Similarity-based Model for Selectional Preferences. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, Prague, Czech Republic.

Kate, R. and Mooney, R.J. (2007). Learning Language Semantics from Ambiguous Supervision. In *Proceedings of the 22nd AAAI Conference on Artificial Intelligence (AAAI)*, 895-900. Vancouver, BC.

Palmer, A., E. Ponvert, J. Baldridge, and C. Smith (2007). A sequencing model for situation entity classification. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, 896-903. Prague, Czech Republic.

Reyle, U., A. Rossdeutscher and H. Kamp (2007). Ups and downs in the theory of temporal reference. *Linguistics and Philosophy*, 30(5):565-635.

Wong, Y.W. and Mooney, R.J. (2007). Learning Synchronous Grammars for Semantic Parsing with Lambda Calculus. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistic (ACL)*, 960-967. Prague, Czech Republic.

2006

Asher, N. (2006). Aspects of Things. In *Philosophical Issues: A Supplement to Nous*, Special number of *Philosophy of Language*, 7:1-20. Blackwell Publishing.

Beavers, J. and A. Koontz-Garboden (2006). A Universal Pronoun in English? *Linguistic Inquiry*, 37:503-513.

Boas, H.C. (2006). From the Field to the Web: Implementing Best-Practice Recommendations in Documentary Linguistics. *Language Resources and Evaluation*, 40(2):153-174.

About the U. of Texas at Austin: The depth, breadth and excellence of UT's research puts it among the top public research institutions in the United States. It has more than 100 research units, some of which, including the Institute for Computational Engineering and Sciences, have been recently created to lead discovery in emerging areas of science. The annual research funding reached $511 million in 2008 (date of the last research report available). Of these, $23 million were awarded to the College of Liberal Arts, of which the Department of Linguistics is a member. Also, more than 400 patents have been awarded to the university since its inception.

Since 1984, more than 40 $1 million-endowed chairs have been created at UT to recruit internationally recognized faculty to accelerate research programs. The faculty at UT is composed of a Nobel laureate, two Pulitzer Prize winners, several MacArthur fellows and hundreds of members of prestigious academic and scientific organizations such as the National Academy of Sciences, the American Academy of Arts and Sciences, the American Philosophical Society and the National Academy of Engineering. This lively faculty has lead to the creation of 36 viable over time start-up companies. The university has one of the largest graduate schools in the nation with more than 10,000 students and more than 170 graduate degree programs. It also is in the top three universities in the number of master's and doctor's degrees awarded annually.

3. Projecte de recerca que es vol desenvolupar durant els primers dos anys d'estada postdoctoral a l'estranger (màxim 10 fulls) / *Research project to be developed during the first two years of postdoctoral abroad (maximum 10 pages)*

3.1 Breu descripció del projecte i dels seus antecedents / *Background and brief description of the research project*
3.2 Objectius, metodologia, pla de treball i bibliografia/referències / *Objectives, methodology, work plan and bibliography/references*
3.3 Impacte previst dels resultats del projecte / *Impact of the project results*

3.1 *Background and brief description of the research project*

Note: I plan to apply for a Beatriu de Pinós grant for a return phase in Catalonia (U. Pompeu Fabra, henceforth UPF). Therefore, the project that follows comprises the two phases (outgoing: 2 years at UT, and return: 1 year at UPF).

The representation of word meaning in formal approaches to semantic theory has traditionally been too limited to account for the richness in descriptive content that words evoke, as this aspect, as well as the interpretive result of combining words, has generally been outside the main concerns of formal semantic theory (Partee 1996; Marconi 1997). Nonetheless, formal semantics has been able to account for a wide range of facts involving the mechanisms of *semantic composition*, that is, the construction of the meaning of a complex expression from the meanings of its parts (Montague 1974), enabling enormous advances in our understanding of linguistic meaning. For many purposes, however, accounting for aspects of meaning beyond the purview of semantic composition is crucial, from practical applications involving human-machine interaction, to the study of concepts from a cognitive perspective. Traditional formal semantic methods therefore need to be supplemented.
Recently, *vector-space* or *distributional* models of meaning (see Turney and Pantel 2010 for an overview) have drawn increasing interest in both the computational linguistics community and the cognitive science community (Sahlgren 2006; Padó and Lapata 2007; Erk and Padó 2008; Andrews et al. 2009; Baroni et al. 2010; Baroni and Lenci 2010). These models provide a very rich, if typically unstructured, representation of a word's meaning. The representation is a vector with hundreds or thousands of dimensions (or *features*), whose values are computed from the co-occurrences of the word in question with other words in a large body of texts, or *corpus.*

Distributional models are feature-based, as are many models in formal semantics and cognition (Pustejovsky 1995; Murphy 2004). They provide a continuous, rather than discrete, representation of meaning, which is compatible with research in cognitive science showing that concepts in the human mind have no clear-cut boundaries, and that they exhibit typicality effects (Rosch 1975 and subsequent work). They also allow for comparisons between words in terms of vector operations such as computation of cosine similarity. As a result, they can model perceived semantic similarity and gradations of similarity between words, something that is hard to achieve with more traditional models. Moreover, they have been shown to be able to reproduce other kinds of human behaviour in different phenomena related to word meaning, such as essay scoring (Landauer and Dumais 1997). Given that nowadays large text corpora are easily available, and given that the resulting models are able to capture many aspects of a word's meaning, these models offer an attractive alternative and complementary view to traditional, symbolic approaches to word meaning.

However, it is not clear whether distributional models will be able to account for aspects of meaning that have been traditionally considered in formal semantics. If they are to be seriously taken as a model of word meaning, they should be able to do so. Given this state of affairs, the goal of the present proposal is to **integrate distributional and corpus-based approaches to word meaning into a theory of semantic composition**. It aims at contributing to **semantic theory** by taking advantage of insights and methods from **computational linguistics** and results from **cognitive science**. Although no psycholinguistic experiments will be carried out at this stage, in the long run the presented

research aims at providing models of word meaning and meaning composition that are adequate from a cognitive point of view. Correspondingly, certain models and datasets developed in cognitive science, particularly in research on the representation of concepts in the human mind, will be taken into account in the computational and theoretical models that will be developed.

3.2 *Objectives, methodology, work plan and bibliography/references*

Specific objectives: We will examine the problem from two different angles: 1) lexical, by looking at regular polysemy (see Work package 1 below); and 2) compositional, by examining noun modification (see Work package 2). Understanding regular polysemy should also shed light on what happens during meaning composition, as we will be looking for contextual clues that help ultimately in resolving the sense alternation. Conversely, analysing noun-modifier composition should shed light on how we get from rather vague words to quite specific meanings. The two phenomena therefore shed complementary light on the general issue we are pursuing, that is, the mechanisms of meaning production and interpretation. We next explain these subprojects in more detail.

Work package 1: Regular polysemy. Many words are *polysemous*, that is, they exhibit more than one meaning or *sense*. With few exceptions (Pustejovsky 1995 being a notable one), the dominant approach both in theoretical and computational semantics to describing sense variation has been a *sense enumeration* approach: For instance, a computational resource such as WordNet (Fellbaum 1998) encodes one sense for *chicken* corresponding to the animal (as in *The chicken slept*) and one referring to its meat (as in *This restaurant serves deep-fried chicken*), among others.
However, many sense alternations are productive, that is, they are not idiosyncratic variations but bear regularities that can easily be reproduced with new words (Apresjan 1974). Consider for instance the case of *chicken*. The alternation between an animal and a meat sense is shared with other nouns, such as *lamb* or *salmon*. Moreover, it is productive: upon hearing sentence (1), we infer that the (invented) noun *wampimuk* refers to an animal; and then sentence (2) is effortlessly interpreted as referring to its meat.

(1) We found a little, hairy wampimuk sleeping behind the tree. (M. Baroni, p.c.)
(2) Wampimuk soup is delicious!

The animal/meat alternation is an instance of a more general mechanism called *grinding* (see Copestake and Briscoe 1995 for discussion and pointers). Because it is a well known alternation in theoretical linguistics, we believe this is a good case study to test distributional approaches to meaning. Our hypothesis is that a regular sense alternation will be signalled by a regularity in the feature representation; therefore, we are specifically interested in analysing whether there are regularities in the features that these models provide for the two senses involved in the regularity for different words. This study requires using *token* (word-in-context) vectors, rather than *type* (word in isolation) vectors; fortunately, there have been several recent approaches to doing so (Erk and Padó 2008 and 2010, Thater et al. 2010, Mitchell and Lapata to appear).

The question of whether a distributional model can represent this kind of regular sense alternation can be subdivided into two separate questions: (1) Given a single lemma (like *chicken*), can a distributional model distinguish its different sense occurrences (e.g. the animal and meat sense)? (2) Are the features that distinguish these senses the same across lemmas in the same semantic class (e.g. the class of nouns denoting animals)? We will study both questions using both supervised (classification) and unsupervised (clustering) models. The answer to question (1) is most likely yes, as the two senses can be expected to occur in suitably different contexts. To be able to say that a distributional model can capture the regularity of the alternation, however, we also need to be able to answer question (2) positively. In the context of question (2), it will be especially interesting to develop techniques for feature selection, that is, to automatically determine features that capture the regularity in the sense alternation across lemmas.

One key research question involved in this subproject is how to build a model that can **generalize to other cases of regular polysemy**. To test this issue, we will test the model on another type of regular polysemy, namely, the alternation between the *stative* and *eventive* readings of deverbal adjectives. These include examples such as *balanced* in English or *cridaner* (vociferous vs. loud-colored) in Catalan (Boleda 2007). In these cases, there is a regular alternation between the truly deverbal meaning of the adjective (e.g., *balanced* as in *the budget was balanced after a long struggle*) and a stative meaning (e.g., *she is truly a balanced person*). For these experiments, we will develop an additional distributional model for Catalan from corpora already available (Reese et al., 2010). Note that almost all of the research on distributional models has been conducted on English, due to the availability of tools and resources for that language, from large corpora and syntactic parsers to datasets from cognitive science; we aim at pursuing this research from a cross-linguistic perspective.
The study of regular polysemy, if successful, will contribute to our larger goal, because it should lead to a model of how language learners pick up on such regularities. This is relevant for cognitive science because regular sense alternations are a core property of natural languages: "the fact of polysemy reveals that it is apparently easier for people to take old words and extend them to new meanings than to invent new words ... [this] is the preferred route even if it results in very complex word meanings". (Murphy 2004: 406). It is also relevant for computational linguistics, mainly for the task of Word Sense Disambiguation (WSD; see Navigli 2009 for a recent survey), the task of identifying the right sense of a word in a given occurrence. The model we will develop should be able to **generalize across lemmata**, such that by modelling regularities in sense alternation, we at least partially overcome the data bottleneck suffered by standard approaches to WSD, at the same time providing a theoretically sounder model.

Work package 2: Semantic composition in the nominal domain. While distributional models are a powerful model for lexical semantics, that is, word meaning, it is currently a pending task to test whether they can account for word meaning in context (the specific use of a given word in a sentence). Researchers are also just beginning to explore how to develop distributional models of complex expressions that can be related in a principled way to the models for their component expressions, that is, that can account for semantic composition (Kintsch 2001; Erk and Padó 2008; Erk and Padó 2010; Lenci and Zamparelli 2010; Mitchell and Lapata to appear).

Most distributional approaches to semantic composition mathematically compose the vectors for each of the parts into a single vector representing the meaning of the whole expression. For instance, for an expression such as *catch a ball* they would sum or multiply the vectors for *catch* and *ball* (Kintsch 2001; Mitchell and Lapata to appear). This is not satisfactory for two reasons: 1) These approaches do not take syntactic structure into account (it is a "bag of words" model), and 2) the resulting vector is simply another point in the vector space, so that the two words are agglutinated into a sort of "catch-ball" word. Erk and Padó (2008) propose a more sophisticated model in which the complex expression contains one vector per word, but each word in the complex expression is modified to account for its meaning in context. Thus, *catch* in the context of *ball* has a different representation than *catch* in the context of *cold*. In more recent work (Erk and Padó 2010), these researchers propose a model, inspired in exemplar-based models in cognitive science, in which only the features of similar sentences are activated.

We will examine the problem of semantic composition tackling examples of noun modification. Noun modification is ideal for this research because it is typically simple in terms of syntax, which allows us to concentrate on the semantic aspects, and because it has been examined in cognitive science from a different angle, as the phenomenon of *conceptual combination* (see Murphy 2004: chapter 12 and pointers there), such that there are specific datasets and phenomena we can computationally model.

2a. Composition of two object-referring expressions. Cross-linguistically, there are several types of modifiers that describe objects related to the entity described by the modified head noun. These include relational adjectives (see example (3)), noun modifiers (example (4)), prepositional phrases (example (5)), and genitives (example (6)):

(3) psychological evidence
(4) world war
(5) agreement by France
(6) Marco's book

It is not clear how distributional models will deal with expressions that denote *tokens* as opposed to describing contexts (*Marco's*, *France).* An important challenge for distributional models is precisely being able to say something about reference, as opposed to just conceptual combination. Reference is something that formal semantic models can handle, so this is again a suitable phenomenon candidate to help us test and further develop distributional approaches to meaning.

In the constructions in examples (3-6), specific meanings with respect to the relationship between the two words arise, often with some kind of "default interpretation" (such as possession for genitives) that can nevertheless be overridden in context. Research in theoretical linguistics (Levi 1978) and in a recent trend in computational linguistics (Girju et al. 2009; Hendrickx et al. 2010) has focused on finding the right relation between a given pair of words. For instance, in Hendrickx et al. (2010) we find the following example:

(7) Instrument-Agency (IA). An agent uses an instrument. Example: *phone operator*

In example (7), *phone* is the Instrument and *operator* the Agent. The problem with this type of approach is twofold: On the one hand, there seems to be no consistent set of relations that account for the relations of all or most examples of nominal modification found in naturally occurring text (one indication for this fact is that every researcher uses a different inventory). On the other, there is much more to the relation between a head and its modifier than the relation itself. This has been convincingly argued in the literature about conceptual combination (Murphy 2004). For instance, Wisniewski (1997) and colleagues have argued that many cases of conceptual combination involve a process of *construal*, in which the meaning of at least one of the components is significantly modified: A plastic truck is a toy, not a real truck; the composition of *truck* and *plastic* alters the meaning of at least the head noun (Partee to appear discusses a similar case, that of *privative adjectives*, from a linguistic perspective).

We will examine whether distributional models can account for the implicit relations between a head noun and its modifier, without necessarily identifying one single relation for the data. We will test the model in the first instance using the paraphrasing dataset for the SemEval-2010 Task #9 (Butnariu et al. 2010), which contains English noun compounds. The task will be to identify from corpus data the adequate paraphrases for a given noun-modifier pair, similarly to what Lapata and Lascarides (2003) did for cases of logical metonymy. We will explore the use of a distributional model for noun compounds akin to the Distributional Memory model (Baroni and Lenci to appear), which joins vector space models and pattern-based extraction. Automatically determining paraphrases in this way is linguistically interesting for the following reasons: (1) It will allow us to test whether there is usually a single paraphrase, or a group of related paraphrases that are appropriate, as Zarcone and Padó (2010) have tested for event-selecting verbs. (2) We will be able to test whether the paraphrases can be categorized into lists of existing semantic relations that other people have proposed. (3) The paraphrasing approach can be expected to work when the relation is conventional and can be determined using world knowledge, but not when the relation is specific to the discourse at hand (see Task 2b).

2b. Discourse and background knowledge effects in noun modification. The relation between the two components in constructions such as (3-6) above are largely left underspecified and vague (Kamp and Partee 1995). We need a theory of how this vagueness is resolved to yield the concrete meanings that these constructions evoke. From preliminary work done by McNally and colleagues (Berndt et al., 2009; McNally 2009; Arsenijevic et al. to appear), it seems that some of the expressions in (3-6) have a particular use, concretely, that relational adjectives (see example in (3) above) are used only when the relation is clear either from the previous context (discourse) or to a broad language community. Prepositional phrases (see (5)), in contrast, can be used to introduce new referents and relations.

A natural test that is in place given this research is that the less material there is in the construction, the more previous background it needs to be able to occur. We will test whether ethnic adjectives (such as *French*) and other types of relational adjectives occur mainly in discourse when there is some sentence that provides the background for the use of the adjective.

We will analyse data from English, Catalan, and German, to analyse this phenomenon from a cross-linguistic perspective. We will develop or adapt additional datasets for these languages including a wider range of constructions (see examples (3-6) above; note that SemEval Task #9 comprises only noun compounds), and track their occurrence in context in natural running text, on corpora already available. The goal will be to test the hypotheses outlined in Arsenijevic et al. (2010) and further hypotheses in a quantitative fashion, using statistical techniques.

Methodology: As can be inferred from the information above, we will combine the traditional methods used in linguistics, that is introspection and construction of positive and negative examples, with computational and quantitative methods. The latter include computational modelling using distributional models, supervised and unsupervised machine learning techniques, and statistical modelling.

We aim at studying some of the phenomena explained above from a cross-linguistic perspective, and to do so in the data-intensive approach defended in this proposal, adequate resources need to be identified or developed. Therefore, we will develop and use computational resources for languages less studied than English, namely German and Catalan. Note that, as stated above, computational and cognitive research is heavily dominated by English, due to the availability of tools and resources for that language, from large corpora and syntactic parsers to datasets in cognitive science.

Work plan: The work plan for the fellowship is as follows (please note that only collaborators external to the outgoing and return host institutions are explicitly included in what follows):

Year 0 (before the start of the fellowship)

1. K. Erk and G. Boleda write a project for the National Science Foundation (Linguistics Program; deadline: 15 January 2011), probably with the involvement of D. Beaver.

Goals to achieve:

✓ recruit additional help to carry out the research related to Work package 1. It will include collecting appropriate corpus data for testing purposes and developing the adequate architecture to perform the computational study.

Year 1

1. Modelling experiments for regular polysemy, *grinding* phenomenon (Work package 1).

- Additional collaborators for this task: M. Baroni (U. Trento), S. Padó (U. Heidelberg).
- Includes the development of the necessary computational architecture, adapting the software available at UT (most notably, software to compute vectors for word meaning in context and S. Padó's software for computing vector spaces) whenever possible.

2. Computational experiments on paraphrasing relation induction for noun compounds (Work package 2a).

- Additional collaborator for this task: S. Padó (U. Heidelberg).

Goals to achieve:

✓ Set up the working architecture that will be the basis for the research programme to be carried out during the fellowship.

✓ Submit one high-impact conference article on the computational modelling of regular polysemy.

✓ Submit one workshop or conference article on the induction of paraphrasing relations for noun compounds.

Year 2

1. Modelling experiments for regular polysemy, generalitzation to the stative/eventive case (Work package 1).

- Additional collaborators for this task: R. Marín (U. Lille).
- Includes the development of datasets for Catalan (partially available) and English (to be developed).

2. Computational experiments for broader noun-modifier constructions (Work package 2a).

- Includes the adaptation of datasets for English.

3. Participation in the follow-up project proposal to OntoSem 2 with L. McNally and colleagues at UPF.

Goals to achieve:

- ✓ Generalize the computational architecture and make it more robust.
- ✓ Build datasets for stative/eventive regular polysemy and release them to the research community.
- ✓ Submit one workshop or conference article about the generalization of the model for regular polysemy.
- ✓ Submit one journal article on the theoretical implications of the computational models for regular polysemy.
- ✓ Submit one high-impact conference article about modelling noun-modifier constructions in English.
- ✓ Prepare a good research framework for the return phase in terms of one funded project.

Year 3

1. Statistical analysis of the effects of discourse and background knowledge in noun modification (Work package 2b).

- Additional collaborators for this task: S. Schulte im Walde (U. Stuttgart), R. Fernández (U. Amsterdam), student assistant.
- Includes the development of the new datasets in Catalan and German with broader constructions (relational adjectives, PP, genitive, morphologically complex nouns).

2. Adapt or develop a theoretical model for the semantic representation of nouns and the semantic composition of noun-modifier constructions.

- Put together all the evidences gathered through the different experiments, and integrate with current research in semantic theory.
- Explore its relation with research in cognitive science (concept representation, conceptual combination, reasoning).

3. Organize the fifth edition of the Quantitative Investigations in Theoretical Linguistics (QITL) workshop, either in Barcelona at UPF or in conjunction with a major international linguistics conference. Co-organisers: to be determined.

- Topic: quantitative, data-intensive approaches to linguistics.

Goals to achieve:

- ✓ Build new datasets for noun-modifier constructions in Catalan and German and release them to the research community.
- ✓ Integrate the results of the different experiments into a consistent theoretical body.
- ✓ Submit one workshop or conference article regarding the development of the new datasets.
- ✓ Submit one journal or high-impact conference article on the role of discourse and background knowledge in the semantic composition of two object-referring expressions.
- ✓ Submit one journal article about the theoretical semantic model.
- ✓ Foster the research community on data-intensive approaches to linguistics through the organization of the QITL workshop.

References

Andrews, M., G. Vigliocco, and D. Vinson (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, 116(3):463-498.

Apresjan, J. D. (1974). Regular Polysemy. *Linguistics*, 142:5-32.

Arsenijevic, B., B. Gehrke, G. Boleda, L. McNally (to appear). Ethnic adjectives are proper adjectives. In *Proceedings of 46th Annual Meeting of the Chicago Linguistic Society*, Chicago, IL, USA.

Asher, N. (in press). *Lexical Meaning in Context*. Cambridge: Cambridge University Press.

Baroni, M. and A. Lenci (2010). Distributional Memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):1-49.

Baroni, M., B. Murphy, E. Barbu, and M. Poesio (2010). Strudel: A corpus-based semantic model based on properties and types. *Cognitive Science*, 34(2):222-254.

Berndt, D., G. Boleda, B. Gehrke, and L. McNally (2009). Nominalizations and nationality expressions: A corpus analysis. Paper presented at the 5th Corpus Linguistics Conference, Liverpool, UK.

Boleda, G. (2007). *Automatic acquisition of semantic classes for adjectives*. Ph.D. thesis, Universitat Pompeu Fabra.

Butnariu, C., S. N. Kim, P. Nakov, D. Ó Séaghdha, S. Szpakowicz, and T. Veale (2010). SemEval-2 Task 9: The Interpretation of Noun Compounds Using Paraphrasing Verbs and Prepositions. In *Proceedings of the 5th International Workshop on Semantic Evaluation* at ACL 2010, Uppsala, Sweden, 39-44.

Copestake, A. and T. Briscoe (1995). Semi-productive polysemy and sense extension. *Journal of Semantics*, 12:15-67.

Erk, K. and S. Padó (2008). A structured vector space model for word meaning in context. In *Proceedings of EMNLP 2008*, Honolulu, HI.

Erk, K. and S. Padó (2010). Exemplar-Based Models for Word Meaning In Context. In *Proceedings of ACL 2010*, Uppsala, Sweden.

Fellbaum, C. (Ed.) (1998). *WordNet: an electronic lexical database*. London: MIT Press.

Gardenförs, P. (2000). *Conceptual Spaces: The Geometry of Thought*. London: MIT Press.

Girju, R., P. Nakov, V. Nastase, S. Szpakowicz, P. Turney, and D. Yuret (2009). Classification of semantic relations between nominals. *Language Resources and Evaluation*, 43(2):105-121.

Hendrickx, I., S. N. Kim, Z. Kozareva, P. Nakov, D. Ó Séaghdha, S. Padó, M. Pennacchiotti, L. Romano, and S. Szpakowicz (2010). SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations between Pairs of Nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation* at ACL 2010, Uppsala, Sweden, 33-38.

Hey, T., S. Tansley, and K. Tolle (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research.

Kamp, H. and B. Partee (1995). Prototype theory and compositionality . *Cognition*, 57(2):129-191.

Kintsch, W. (2001). Predication. *Cognitive Science*, 25:173-202.

Landauer, T. and S. Dumais (1997). A solution to Platos problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211-240.

Lapata, M. and A. Lascarides (2003). A Probabilistic Account of Logical Metonymy. *Computational Linguistics,* 29(2):263-317.

Lenci, A. and R. Zamparelli (Eds.) (2010). *Proceedings of the ESSLLI 2010 Workshop on Compositionality and Distributional Semantic Models*, Copenhagen, Denmark.

Levi, J. N. (1978). *The Syntax and semantics of complex nominals*. New York: Academic Press.

Liberman, M. (2010). The Future of Computational Linguistics: or, What Would Antonio Zampolli Do? Antonio Zampolli Prize Talk at LREC 2010, Valletta, Malta.

Marconi, D. (1997). *Lexical competence*. Cambridge, MA: MIT Press.

McNally, L. (2009). How much vagueness needs resolving? Invited talk at the ESSLLI 2009 Workshop on Vagueness in Communication.

Mitchell, J. and M. Lapata (to appear). Composition in Distributional Models of Semantics. *Cognitive Science*.

Montague, R. (1974). English as a formal language. In Thomason, R. H. (Ed.), *Formal philosophy: Selected Papers of Richard Montague*, chapter 6: 188-221. New Haven: Yale University Press.

Murphy, G. L. (2004). *The big book of concepts*. Cambridge, MA: MIT press. Paperback edition.
Navigli, R. (2009). Word Sense Disambiguation: A Survey. *ACM Computing Surveys*, 41(2):1-69.
Padó, S. and Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2), 161-199.
Partee, B. H. (1996). The development of formal semantics in linguistic theory. In Lappin, S. (Ed.), *The Handbook of Contemporary Semantic Theory*, pages 11-38. Oxford: Blackwell.
Partee, B. H. (to appear). Privative adjectives: subsective plus coercion. To appear in Zimmermann, T. E. (Ed.), *Studies in Presupposition*.
Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge, MA: MIT Press.
Reese, S., G. Boleda, M. Cuadros, L. Padró, G. Rigau. 2010. Wikicorpus: A word-sense disambiguated multilingual Wikipedia corpus. In *Proceedings of LREC 2010*, Valletta, Malta.
Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104:192-233.
Sahlgren, M. (2006). *The Word-Space Model*. Ph.D. thesis, Stockholm University.
Thater, S. H. Fürstenau, and M. Pinkal (2010). Contextualizing Semantic Representations Using Syntactically Enriched Vector Models. In *Proceedings of ACL 2010*, Uppsala, Sweden.
Turney, P. D. and P. Pantel (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141-188.
Wisniewski, E. J. (1997). When concepts combine. *Psychonomic Bulletin & Review*, 4:167-183.
Zarcone, A. and S. Padó (2010). Implicit Events for Event/Entity-Ambiguous Nouns. Talk at the Workshop on Zugänglichkeit impliziter Ereignisse, Tübingen, Germany.

3.3 *Impact of the project results*

Taking all the above into account, this project will have an impact along three different dimensions of research.

1) Theoretical: It will bring together the complementary insights provided by formal semantics and distributional models to help develop a sounder theory of word meaning and meaning composition, and to contrast these results with results obtained in research in cognitive science, thus **advancing the state of the art knowledge on the mechanisms of meaning production and interpretation**. This will be achieved by bringing together insights about meaning from three fields which have not been sufficiently integrated, namely, theoretical linguistics, computational linguistics, and cognitive science. For semantic theory, this will mean overcoming the fact that in formal semantics most of lexical semantics has been relegated to a "separate empirical domain" (Partee 1996: 34), not addressing issues of interpretation of concern to, e.g., cognitive scientists and computational linguists. For these two fields, accounting for word meaning is a pressing need; in computational linguistics, it has long been recognized that the lexicon, and in particular lexical semantics, is a bottleneck to building useful applications; in cognitive science, word meaning is the window to exploring concepts and conceptual representations, a major building block of general reasoning abilities, that what makes humans unique. Therefore, accounting for word meaning and meaning composition in a manner that is empirically adequate and robust is central to all three disciplines. We believe that the approach proposed holds promise of shedding some light on more general aspects of human cognition and interaction. Furthermore, note that although the applicant will not carry out psycholinguistic experiments in this phase, the fellowship will allow her to prepare the ground for collaborations involving neuro- and psycholinguistic experiments to test the hypotheses and models elaborated in the present phase. The continued contact with Brain and Cognition groups at UT, U. Trento and UPF will make this possible.

2) Methodological: It will contribute to the development of adequate procedures and resources (from language resources to appropriate statistical and machine learning

techniques) to explore these issues on a large scale, thus helping establish standards for empirical, quantitative, or data-intensive linguistics, an emerging approach to the study of language. To this respect, note that the project tackles questions about semantic theory with a novel methodology, **using statistical and computational methods to gain insight into the nature of human languages**.
Traditional methods in generative approaches to linguistics (introspection, self-construction of positive and negative examples) need to be supplemented when it comes to phenomena with no clear boundaries, such as phenomena having to do with lexical semantics, about which it is difficult to have intuitions. Data-intensive approaches to language hold promise of offering new insights not afforded by traditional linguistic methodologies, by systematically testing hypotheses at a large scale and controlling for sources of variation with statistical techniques. The development of computational techniques and the hardware itself, and more importantly, the availability of massive amounts of language data, make it for a ripe time to explore long-standing puzzles such as polysemy from a linguistic perspective with data-intensive methods. This methodology has been the focus of computational linguistics for the last 15 years, and it is also in line with other sciences, which are actively engaged in the emergence of the so-called "Fourth paradigm" in science (Hey et al. 2009), that is, data-intensive scientific research. While Europe is still largely lagging behind, in the US there are some steps being taken to bring this development to linguistics (Liberman 2010), and an active focus is precisely the University of Texas at Austin, the host institution of the present proposal. Moreover, the **integration of cognitive results and methods in linguistics**, which is not common, as it requires highly interdisciplinary training and multidisciplinary collaboration, will also have an impact on the study of language. Note that these two methodological innovations will not only impact semantic theory, but linguistics in more general terms, as many of the techniques and experimental designs can be generalized to other areas, such as syntax or morphology.

3) Social. World languages are in different sociolinguistic situations due to a variety of factors, including historic development, size of the speaker community, and the different official statuses of the languages in question. The stronger ones present an acceptable degree of computational development and linguistic tools (including corpora, lexica, experimental datasets, and processing tools), whereas the weaker ones have fewer –if any– resources. Moreover, as mentioned above, research in distributional models has been carried out almost exclusively for English, due to the availability of resources and datasets for this language. By building datasets and models for Catalan and German (in addition to English), and making them available to the research community, this project supports linguistic diversity.

Moreover, the mobility will strengthen not only the applicant's research skills, but also those of research groups in Spain, Germany, Italy, France, and the Netherlands with whom the applicant has ongoing collaborations:

- U. Pompeu Fabra, in particular L. McNally's OntoSem 2 and REDISEM projects;
- U. Stuttgart, in particular S. Padó's project SFB 732-D6 *Lexical-semantic factors in event interpretation* (granted) and Sabine Schulte im Walde's project *Distributional Approaches to Semantic Relatedness* (pending evaluation; both related to the topic of Work package 2a), and the TransCoop proposal between S. Padó and K. Erk, related to Work package 1;
- U. Heidelberg, where S. Padó has moved to occupy a professorship as of October 2010;
- U. Trento, through the REDISEM project between L. McNally, G. Boleda, M. Baroni, and R. Zamparelli, which also attempts at bringing together distributional and formal approaches to semantics in two phenomena complementary to the ones studied in this proposal;
- U. Lille, in particular R. Marín's project *Analyse sémantique et codification lexicale des nominalisations*, related to Work package 2a, and a future collaboration related to Work package 1;
- U. Amsterdam, through the collaboration with R. Fernández on Work package 2b.

More generally, it will reinforce the research community on semantic theory, data-intensive linguistics, computational linguistics, and cognitive science.

4. **Aspectes ètics (màxim 2 fulls) / *Ethical issues (maximum 2 pages)***

3.1 Indiqueu si el projecte de recerca que es vol desenvolupar inclou algun d'aquests aspectes ètics i, si el projecte n'inclou algun, expliqueu-ne breument els motius / *Indicate whether the research project to be developed includes some of these ethical issues and, If the project includes some of them, explain briefly why:*

	Si/*Yes*	No
• Investigació amb embrions o cèl·lules embrionàries humanes / *Research on human embryos or embryonic stem cells*	☐	☐
• Investigació que involucra teixits o cèl·lules fetals humanes / *Research that involves human foetal tissues or cells*	☐	☐
• Investigació amb persones menors d'edat o amb persones incapaces que no poden donar el seu consentiment[1] / *Research with minors or incapable people who cant not give consent*[1]	☐	☐
• Investigació utilitzant tècniques invasives en els pacients[1] / *Research using invasive techniques on patients* [1]	☐	☐
• Investigació amb voluntaris adults sans[1] / *Research using adult healthy volunteers*[1]	☐	☐
• Investigació amb material genètic o mostres biològiques humanes / *Research with human genetic material or biological samples*	☐	☐
• Investigació que involucra recollida de dades humanes[1] / *Research that involves human data collection*[1]	☐	☐
• Investigació amb animals, animals de granja modificats genèticament o de laboratori / *Research with animals or genetically-modified farm or laboratory animals*	☐	☐

1). En els casos de que el projecte presentat inclogui aquest tipus d'investigació, també caldrà especificar si existeix algun tipus de remuneració o de compensació per als subjectes participants. En el moment de presentar la sol·licitud, també caldrà adjuntar el model d'informació i de consentiment que rebran els participants / *In the event that the project presented include this type of research, also must specify whether there is any kind of remuneration or compensation for participating subjects. When submitting the proposal, you must also attach the information and consent model that participants will receive*