MINISTERIO
DE ECONOMÍA
Y COMPETITIVIDAD

SECRETARÍA DE ESTADO
DE INVESTIGACIÓN
DESARROLLO E INNOVACIÓN

SECRETARÍA GENERAL
DE CIENCIA, TECNOLOGÍA
E INNOVACIÓN

DIRECCIÓN GENERAL
DE INVESTIGACIÓN
CIENTÍFICA Y TÉCNICA

SUBDIRECCIÓN GENERAL
DE RECURSOS HUMANOS
PARA LA INVESTIGACIÓN

AYUDAS RAMÓN Y CAJAL

MEMORIA DE LA TRAYECTORIA INVESTIGADORA Y LA LÍNEA DE INVESTIGACIÓN PRINCIPAL QUE HA DESARROLLADO (*SUMMARY OF THE RESEARCH CAREER OF THE CANDIDATE AND THE MAIN RESEARCH LINE THAT SHE/HE HAS CARRIED OUT)*

Esta memoria debe rellenarse preferiblemente en inglés - *Summary to be completed preferably in English*

INVESTIGADOR SOLICITANTE / *RESEARCHER APPLICANT*: Gemma Boleda

PALABRAS CLAVE / *KEY WORDS:* *computational semantics, semantic theory, distributional semantics, formal semantics, lexical semantics, reference*

RESUMEN (aprox. 300 palabras) / SUMARY (approx. 300 words):

My primary research interest lies in understanding how meaning works in natural languages. I approach this research goal from an interdisciplinary perspective, using computational linguistic methods to answer theoretical semantic questions. I have worked on lexical semantics and semantic composition, as well as on building resources that enable data-intensive research on language. In the next phase of my career, I plan to focus on the major enterprise of using the empirical results and conceptual findings of my research to propose a comprehensive semantic theory that encompasses lexical and structural aspects of meaning.

My research has a clear impact, with 582 citations, an h-index of 12, and 16 articles in the top venues in Computational Linguistics according to Google Scholar, as well as four articles in journals indexed in the ISI Web of Knowledge. My intense international activity includes over four years of working experience in five universities in Germany, Italy, and the US, and participation in five international projects including one European Network of Excellence and one project funded by DARPA (US). The international community, in turn, recognizes me as an established researcher in the field, trusting me for instance to act as co-editor for a special issue in a top journal (Computational Linguistics) and to be on the Editorial Board of the Linguistic Issues in Language Technology journal, associated with the Linguistic Society of America. In 2015-2016, I will give four invited talks in international workshops.

Throughout my career, I have shown independence in my research as well as strong leadership, team work, and coordinating skills, for instance by acting as the PI of a Marie Curie grant, collaborating with a broad variety of researchers in different disciplines, mentoring students, and organizing three international workshops. Finally, I have proven my ability to obtain competitive funding, including three prestigious post-doctoral contracts in the Juan de la Cierva, Beatriu de Pinós, and Marie Curie programs.

DESTACAR HASTA UN MÁXIMO DE LAS CINCO APORTACIONES MÁS RELEVANTES ENTRE LAS RECOGIDAS EN EL CURRÍCULUM VITAE

Highlights a maximun of 5 relevant achievements from the ones includes in the cv

Also see Google Scholar profile (https://scholar.google.com/citations?user=NFJ9kUEAAAAJ).

1) **Publication**: **Boleda**, G.; S. Schulte im Walde; T. Badia. Modeling regular polysemy: A study in the semantic classification of Catalan adjectives. *Computational Linguistics*. 38 - 3, pp. 575 - 616. MIT Press, 2012.

   Relevance: Article in the ***top ranked*** *Computational Linguistics journal in the Linguistics category according to the Journal Citation Reports of the ISI Web of Knowledge (henceforth, JCR), with an impact factor of 0.940 in 2012 (****first quartile****, journal 21 of 166 in the Linguistics category in terms of impact factor).*

2) **Publication**: **Boleda**, G.; E. M. Vecchi; M. Cornudella; L. McNally. First-order vs. higher-order modification in distributional semantics. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1223 - 1233. ACL, 2012. Acceptance rate: 24%.[1]

   Relevance: Article in the 2nd ***top venue*** *in Computational Linguistics according to Google Scholar.*[2]

3) **Publication**: Corral, Á., G. **Boleda**, R. Ferrer-i-Cancho (2015). Zipf's Law for Word Frequencies: Word Forms versus Lemmas in Long Texts. PLoS ONE 10(7):doi:10.1371/journal.pone.0129031.

   Relevance: Interdisciplinary research (Linguistics and Physics). General science journal in the ***first quartile*** *of its category (Science, Multidisciplinary), with an impact factor of 3.534 according to JCR.*

4) **Special Issue co-editor**: **Boleda**, G.; A. Herbelot (Eds.). Special Issue on Formal Distributional Semantics. *Computational Linguistics*. In preparation, to be published in the fall of 2016.

   Relevance: As mentioned above, Computational Linguistics *is the top ranked journal of Computational Linguistics according to JCR. Special issues follow a competitive reviewing process, and only respected, senior members of the community with a strong proposal are accepted as editors of special issues for this journal.*

5) **Project Principal Investigator**: "LOVe: Linking Objects to Vectors in distributional semantics: A framework to anchor corpus-based meaning representations to the external world" (EU, Marie Skłodowska-Curie project 655577, H2020-MSCA-IF-2014). Dates: 2015-2017. Total granted amount: 180,277.20€.

   Relevance: Highly competitive European Union grant in the Horizon 2020 framework that showcases my ability to obtain external funding.

[1] Source: http://www.aclweb.org/aclwiki/index.php?title=Conference_acceptance_rates.

[2] https://scholar.google.es/citations?view_op=top_venues&vq=eng_computationallinguistics.

DESARROLLAR LA TRAYECTORIA INVESTIGADORA ASÍ COMO LA LÍNEA DE INVESTIGACIÓN PRINCIPAL QUE HA DESARROLLADO.

Extended detail of the research career of the candidate and the main research line that he/she has carried out.

(El tamaño máximo del fichero será de 4 Mb / The maximun file size will be 4 MB)

1. Career

I discovered what would become my passion, Linguistics, during my undergraduate studies in Spanish Philology at the Universidad Autónoma de Barcelona, where I received a solid training in this discipline. Towards the end of the degree I became interested in Computational Linguistics (also known as Natural Language Processing), an interdisciplinary field involving Linguistics and Artificial Intelligence with a strong empirical focus, which allows hypotheses to be tested through well-defined computer simulation experiments, statistical analyses, and ultimately applications such as Machine Translation. Since there was little Computational Linguistics research being undertaken at my university, I decided to further pursue undergraduate studies at the Department of Linguistic Information Processing (University of Cologne, Germany) with the aid of an Erasmus fellowship. There I had my first hands-on experience as a researcher, since I was hired as a student assistant to work on a Machine Translation project in the department. Back in Barcelona, while finishing my degree (with an End of Degree Award by the university and a Special Mention in the national degree awards), I worked at Pompeu Fabra University as a student assistant on a corpus linguistics project for educational purposes, and later at the Artificial Intelligence Research Institute (IIIA, CSIC) through a three-month Introduction to Research fellowship awarded by the CSIC.

I then moved to Pompeu Fabra University to pursue a PhD in Computational Linguistics under the supervision of Toni Badia. In Badia's group, GLiCom, I enjoyed a rich research environment, with a dozen other PhD students over time, seminars and reading groups, and regular interaction with other professors in the department. A particularly fruitful collaboration with Louise McNally started in that period and continues to date. I also had the opportunity to participate in two national and three international projects headed by T. Badia and L. McNally, both as a researcher and also contributing to the write-up of the project proposals, something PhD students rarely get the chance to do. In this phase, I was trained as a computational linguist, not only through the thesis but also participating in the development of language processing tools and resources such as a part-of-speech tagger, a syntactic parser, and several corpora. I also started my international activity, collaborating with researchers in the international projects of the group, presenting my work in relevant international forums, and acting as a reviewer for the most relevant conferences in the field (details in Sections 3 and 4 below). I also obtained competitive funding to carry out two research visits at the CoLi (University of Saarland, Germany), one of the top research centers in Computational Linguistics in Europe. During the second visit I met Sabine Schulte im Walde, an internationally recognized researcher in computational semantics, who accepted to act as my thesis co-supervisor. Finally, in that period I also gained teaching experience at Pompeu Fabra University as a teaching assistant.

Upon completion of my PhD, I obtained funding from the *Juan de la Cierva* program of the Spanish government for a three-year post-doc in the Natural Language Processing group (GPLN) of the Department of Software at the Polytechnic University of Catalonia (UPC). Since my training had been mainly in Humanities departments, working at the GPLN was very useful for me to further develop my mathematical and computational skills. During that phase I collaborated with my mentor, Lluís Padró, and his students in the development of the open source language processing tool FreeLing, I led the construction of freely available corpora for English, Spanish, and Catalan, and I participated in two national projects involving the Universities of Barcelona and the Basque country, as well as the EU Network of Excellence PASCAL 2. I also kept collaborating with Louise McNally, and was involved in two national projects led by her. In this period, I started taking on more senior roles, such as co-organizing international scientific events and mentoring students: Daniel Berndt (student assistant), Samuel Reese (Master's thesis), and Cristina Sánchez-Marco (thesis project; co-supervisor J. M. Fontana). Finally, I carried out a five-month research visit at IMS (Institute for Natural Language Processing, Stuttgart University, Germany), also one of the top European research centers in the field, with competitive funding from the European Network of Excellence PASCAL 2 and the SFB 732 project of Stuttgart University. There I worked with Sebastian Padó, Sabine Schulte im Walde, and their research groups.

After further working for ten months as a researcher in a project of the prestigious EXPLORA program headed by Louise McNally at the Pompeu Fabra University (within which I co-supervised another Master's thesis, Miquel Cornudella's), I moved to the Department of Linguistics of the University of Texas at Austin (USA) with funding from a *Beatriu de Pinós* fellowship from the Catalan government (co-funded by the Marie Curie programme of the European Union). There I worked as a post-doctoral researcher and later I was trusted by the department chair to take on a lecturing role. With respect to my research field, The University of Texas at Austin is quite unique in the US, as it features a very strong profile in both Linguistics and Computer Science. My mentor, Katrin Erk, and Ray Mooney, the head of the Artificial Intelligence Lab in the Computer Science department, were the co-PIs of a large DARPA (Defense Advanced Research Projects Agency) project I participated in. I also regularly interacted with world-renowned linguists such as Hans Kamp, David Beaver, John Beavers, and Stephen Wechsler, as well as

DESARROLLAR LA TRAYECTORIA INVESTIGADORA ASÍ COMO LA LÍNEA DE INVESTIGACIÓN PRINCIPAL QUE HA DESARROLLADO.

Extended detail of the research career of the candidate and the main research line that he/she has carried out.

(El tamaño máximo del fichero será de 4 Mb / The maximun file size will be 4 MB)

their students, in classes and reading groups. After 26 months, I returned for one year to Pompeu Fabra University in the second phase of the *Beatriu de Pinós* fellowship, during which I obtained a highly competitive Marie Curie Marie Sklodowska-Curie Individual Fellowship (H2020-MSCA-IF-2014-655577, 180,277.20€) to carry out a two-year project on theoretical and computational semantics at the University of Trento, where I work since June 2015. In this latter phase of my career I have gained wide recognition from the scientific community, as shown for instance by the fact that I have been invited to give four talks at international workshops (in Spain, Israel, the US, and Japan). Also, I am regularly asked to carry out high-responsibility scientific evaluation tasks (such as reviewing for the most important journals in the field, like *Computational Linguistics*, *Artificial Intelligence*, *Natural Language Engineering*; evaluating projects for research agencies in Argentina and Poland; and co-editing a special issue in *Computational Linguistics*) as well as prominent roles in the organization of international scientific events (area co-chair of ACL 2016; program co-chair of *SEM 2015; area co-chair of *SEM 2013; local co-chair of ESSLLI 2015; see Sections 3-5 for more information about these events).

2. Research

My primary research interest lies in understanding how *meaning* works in natural language. I have worked on **lexical semantics** (the meaning of words) and on **semantic composition** (how the meaning of words is combined to yield the meaning of phrases), and my current focus is on the broader implications of my research for **semantic theory**. As mentioned above, I approach these topics from a **computational** and **quantitative** perspective, as I believe that this perspective fruitfully complements traditional, symbolic and qualitative approaches to semantics in Linguistics. I started out working in the semantic tradition known as **formal semantics** (Montague 1974 and subsequent work), which provides a rigorous approach to semantic composition by using tools from logic and set theory. This approach has enabled great progress in our understanding of various semantic phenomena. Its representations give detailed information about logical or *structural* aspects of language, such as the number of distinct events and individuals identifiable in a sentence and the relationships between them. They also provide a suitable frame for predicting inference relations. Despite these successes, though, formal approaches have left the semantics of content words such as *book*, *run*, or *red* largely unexamined (with some exceptions such as Pustejovsky 1995 and Asher 2011). Indeed, lexical semantics is pervasively affected by phenomena such as ambiguity and vagueness, which demand approaches that naturally encompass a notion of degree in their representations. First-order logic, using atomic symbols, struggles with these properties of natural languages.

For this reason, I have also embraced **distributional semantics**, an inductive approach to meaning currently very popular in Computational Linguistics and Cognitive Science (see Turney and Pantel 2010 for an overview). Distributional representations encode lexical meaning as a function of the contexts in which words occur, where the context for a word can be defined for instance as the set of words in the same sentence, a document it appears in, or in more sophisticated ways that take syntax into account. Distributional representations encode very rich information in a graded manner. Accordingly, distributional models successfully account for many lexical semantic phenomena: For instance, they can replicate human behavior with regard to word similarity (*string* and *cord* are similar, *professor* and *cucumber* are not), priming (humans respond faster to *doctor* if they have read *hospital* before), or synonymy identification (*pinnacle* is a synonym of *zenith*, *outset* is not; see Landauer and Dumais 1997 and subsequent work). The reason distributional semantics is so successful is that its representations are induced from linguistic data naturally produced by humans, mainly collections of text drawn from the internet and other sources. Two recent advances in the field, in which I have been directly involved, are the following: (1) *Compositional* distributional semantics, seeking to combine the meaning representations of words into semantic representations for phrases or even sentences (Mitchell and Lapata 2010, Coecke et al. 2011, Socher et al. 2013, Baroni et al. 2014, refs. 9, 10, and 12 in the CV); and (2) the introduction of *perceptual* information, in particular visual information extracted from images (Feng and Lapata 2010, Silberer et al. 2013, ref. 11 in the CV). Here, recent advances in computer vision allow us to use visual features as contexts, providing better, perceptually grounded representations for words.

Distributional semantics has been very successful in practical applications; for example, its basic principle lies at the heart of Information Retrieval engines (Baeza-Yates and Ribeiro-Neto 2011). However, there is reason to think that distributional semantics is not only a useful engineering device, but that it also has deeper philosophical implications for the study of meaning (Erk 2013, ref. 2 in the CV); in fact, its “meaning as use” approach has a clear antecedent in Wittgenstein's *Philosophical Investigations*. Formal semantics, in turn, is based on a realist framework by which meaning is “out in the world” (Portner 2005, Section 1.2). Given that the theoretical positions and the relative strengths of these two approaches are complementary, my specific goal is to **integrate distributional and formal approaches to semantics**. I next outline my main, interrelated research lines.

DESARROLLAR LA TRAYECTORIA INVESTIGADORA ASÍ COMO LA LÍNEA DE INVESTIGACIÓN PRINCIPAL QUE HA DESARROLLADO.

Extended detail of the research career of the candidate and the main research line that he/she has carried out.

(El tamaño máximo del fichero será de 4 Mb / The maximun file size will be 4 MB)

Lexical semantics. In my PhD thesis, I focused on the semantic classification of adjectives. I defined a new classification at the morphology-semantics and syntax-semantics interfaces, and tested it through both human annotation and computational modeling experiments that applied Machine Learning to distributional semantic representations. These experiments yielded further insight into the linguistic properties of each adjective class. A special focus of the thesis was regular polysemy in adjectives, since many adjectives exhibit different senses that fall into different classes in the proposed classification. I later continued working on regular polysemy, this time examining nouns (ref. 14 in the publication list in the CV). I co-authored eight international publications related to my thesis (refs. 13, 19, 22, 24, 30, 32, 34, 42 in the CV), the most relevant one being an article in a top journal in the field, *Computational Linguistics*, ranked 21/166 in the Linguistics category and with an impact factor of 0.950 according to JCR (ref. 13). I have also explored other topics related to lexical semantics and the automatic induction of linguistic knowledge, such as the semantic classification of verbs (ref. 27 in the CV), their argument structure (ref. 33), and the identification of hypernymy (ref. 5). This research has been published in another top journal in the field, *Language Resources and Evaluation,* also indexed in the JCR, as well as top conferences (EMNLP, COLING) – see Section 3 below for information about these venues.

Semantic composition. I have worked on adjective modification from different angles: From a purely formal semantics perspective (refs. 6, 42), with statistical modeling (refs. 40), and combining formal and distributional approaches to semantic composition (refs. 1, 9, 12). I have done extensive research on adjective-noun composition for one of the adjective classes examined in the thesis, that of relational adjectives (e.g., *pulmonary*), and more recently on a subclass of those, ethnic adjectives (e.g., *French).* Relational adjectives resist a set-theoretic treatment, and it has even been questioned that they are fully-fledged adjectives (Fábregas 2007, Alexiadou and Stavrou 2011). Our proposal is that relational adjectives are proper adjectives, and are best analyzed as intersective properties of kinds; ethnic adjectives, furthermore, invoke a default *Origin* relation that can be overridden in context. The most representative publication in this line of research has received 119 citations to date according to Google Scholar (ref. 42), and another one in a Springer volume (ref. 40).

More recently, I have studied modification by intensional adjectives *(alleged),* compared to standard, non-intensional ones (*big*). Formal semantics assigns a fundamentally different semantic analysis to the two types of adjectives. Surprisingly, our experiments using distributional approaches to composition reveal no differences between the two on a number of measures, and suggest that the typicality of the attribute denoted by the adjective for the nominal concept is a relevant parameter when dealing with adjective modification – one that has been overlooked so far. This computational studies led to theoretical insight regarding conceptual vs. referential effects in modification (see next paragraph). This research has been published, a.o., in the 2nd top venue in the field according to Google Scholar (EMNLP; ref. 9) and a forthcoming Springer volume (ref. 1).

Semantic theory. In my recent research, I aim at abstracting away from the specific phenomena examined and work towards the broader implications of my findings for a general theory of meaning, and this is the enterprise I plan to focus on in the next phase of my career. First, as mentioned above, I have developed perceptually grounded models, by including not only textual but also visual information in distributional models, automatically extracted from images. This addresses a problem that plagues many theoretical and computational approaches to semantics, which rely on other symbols (such as semantic primitives or related words) to represent meaning. The problem is that those symbols also need an interpretation, in a loop that only ends if relevant aspects of meaning can ultimately be grounded in non-linguistic information (Harnad 1990). This research has yielded an article in the top venue in Computational Linguistics (ACL; ref. 11; see Section 3 for more information) which has already been cited 65 times.[3] I am currently working on the bold task of extending the grounding of distributional semantics to actual reference in the external world. My hypothesis is that distributional semantics yields concept-like representations for linguistic expressions that can be anchored in their actual referents in the world. This is work in progress; an initial experiment mapping distributional representations for proper nouns to entities in a database has been published at EMNLP (ref. 3). More generally, I am working on a **theoretical framework that encompasses conceptual and referential aspects of meaning**, comparing distributional with primitive-based approaches (ref. 2) and exploring the combination of formal and distributional semantics (refs. 1, 3, 6). The latter topic is also addressed in the special issue of *Computational Linguistics* I am currently co-editing.

Creation, validation, and promotion of linguistic data. The sort of research I am engaged in relies on large data sources, both language resources, such as corpora or machine-readable lexica, and human annotated data. As mentioned above, I am actively involved in the creation of language resources, especially for Catalan and Spanish,

[3] Source: Google Scholar, http://scholar.google.es/citations?user=NFJ9kUEAAAAJ&hl=en, accessed January 19th 2016.

DESARROLLAR LA TRAYECTORIA INVESTIGADORA ASÍ COMO LA LÍNEA DE INVESTIGACIÓN PRINCIPAL QUE HA DESARROLLADO.

Extended detail of the research career of the candidate and the main research line that he/she has carried out.

(El tamaño máximo del fichero será de 4 Mb / The maximun file size will be 4 MB)

languages with scarce available resources. I strive to make them available to students and researchers, such that they can easily access large amounts of attested linguistic data, contributing to a better empirical footing of linguistic research. These resources include a Constraint Grammar part-of-speech tagger and shallow parser for Catalan and Spanish, annotated corpora and open source part-of-speech taggers for Old Spanish and Old Catalan, the open source suite of language analyzers FreeLing, a freely available Wikipedia corpus for Catalan, Spanish, and English, a Catalan Web corpus with a flexible interface, and a corpus of aligned translations for several European languages built for educational purposes (refs. 15, 17, 19, 20, 23, 25, 35, 36, 37, 38, 39, 41). Given that Computational Linguistics heavily relies on annotations made by human subjects, I have also worked on the assessment of the quality of human annotations for linguistic phenomena, proposing a method to quantify agreement across a large number of judges (published in a Springer journal article, ref. 22) and co-organizing a COLING workshop on the topic to encourage research in this area (ref. 45). I have also collaborated in or led the development of five freely available datasets annotated for semantic phenomena.[4]

I am also interested in enabling other scientific disciplines to carry out research on language using Computational Linguistic / Natural Language Processing tools and techniques, again with the aim of broadening the empirical study of language as much as possible. For instance, I am engaged in a collaboration with researchers in Physics to examine the dynamic properties of word distributions and their relationship to Zipf's law, the best known statistical law for language. Thanks to this collaboration, my co-authors and me can examine statistics not only for raw word forms, as is standard in the field, but also for lemmas. This interdisciplinary research has yielded two high-impact publications: one in PLoS ONE, a general science journal is in the first quartile of its category in JCR (Science, Multidisciplinary) with an impact factor of 3.534 (ref. 4), and one a first-quartile Physics journal with an impact factor of 4.063 according to JCR (ref. 8). I also strive to promote the social use of computational tools and language resources more widely, through the publication of articles in general audience journals and the organization of events (see Section 5 for more information).

3. Contributions (*Aportaciones*).

My contributions have a clear scientific impact, as follows. With 49 publications so far, I have an **h-index of 12** and **582 citations** (source: Google Scholar, see URL provided in footnote 3). I have published **four** articles in journals indexed in the ISI Web of Knowledge Journal Citation Reports (JCR; references 4, 8, 13, and 27 in the CV publication list; another JCR-indexed article will be published in 2016). Three of these articles are in the **first quartile** of the respective categories, including a General Science journal (*PLoS ONE*). I have also published two articles in Springer volumes (refs. 1 and 40) and co-edited three proceedings books, including one international conference proceedings book. The workshop proceedings I have published (one at IWCS and one at COLING; refs. 44 and 45) have high scientific relevance because workshops at conferences, like articles, follow a competitive reviewing process: Workshop submissions are evaluated by the conference organizers, and only solid proposals with organizers that are respected members of the community go through.

The scientific quality of my publications is further attested in Table 1 below, which shows that **one third** of them (16/49) were published in **the top 10 venues** in Computational Linguistics according to Google Scholar (see URL in footnote 2). I have also contributed to the two main conferences specializing in computational semantics, IWCS and *SEM, with three further articles (refs. 9, 10, 14). Note that, like other fields related to Computer Science, Computational Linguistics is primarily conference-based, such that many of its top venues are conferences (see rows 1-5 and 9 in Table 1). The acceptance rates of these conferences are in general extremely low, around 20-30% for the top ones (see URL in footnote 1). Also note that most of my publications are co-authored; again, this is the norm in my field, which is very interdisciplinary and computationally intensive.

Rank	Venue	Pubs.	Ref. in CV
1.	Meeting of the Association for Computational Linguistics (ACL)	1	11
2.	Conference on Empirical Methods in Natural Language Processing (EMNLP)	4	3, 12, 24, 33
3.	North American Chapter of the Association for Computational Linguistics (NAACL)		
4.	International Conference on Language Resources and Evaluation (LREC)	6	16-20, 37
5.	International Conference on Computational Linguistics (COLING)	3	5, 32, 5

[4] See http://gboleda.utcompling.com/resources. The datasets are available under a Creative Commons BY-SA license.

6.	arXiv Computation and Language (cs.CL)		
7.	*Computer Speech & Language*		
8.	*Computational Linguistics*	1	13
9.	Conference on Computational Natural Language Learning		
10.	*Language Resources and Evaluation*	1	27

Table 1. Publications in the top 10 venues in Computational Linguistics according to Google Scholar.

4. Participation in international activity (*Participación en actividad internacional*).

From the early experience at the University of Cologne, my research career has been markedly international, which has allowed me to build a solid research network encompassing Europe and the US. I have worked at three universities in Germany (Cologne, Saarland, and Stuttgart), one in Italy (Trento), and one in the United States (The University of Texas at Austin) at the undergraduate, PhD, and post-doc levels, for a total of over four years of international experience. Furthermore, I have obtained competitive funding from national and international bodies (Catalan, Spanish, and German governments, European Union) to fund these experiences at universities abroad.

I have participated as a researcher in two European projects (METIS-II, eTitle) and one European Network of Excellence (PASCAL 2), as well as one US DARPA project (*Statistical Relational Learning and Script Induction for Textual Inference*), all of them very large projects with over 1,000,000 EUR of funding each. Another relevant international experience was the participation in two Spanish Ministry of Education and Science international cooperation grants with the universities Pompeu Fabra, Osnabrück and Lille. I collaborated in the writing of proposals for international projects from early on, and I have recently obtained European funding as the PI of a Marie Skłodowska-Curie grant.

As a result of this intense international activity, almost 40% of my scientific production (19/49) has been written in collaboration with colleagues working abroad, especially in Germany, Italy, and the US. My work has been disseminated in dozens of contributed presentations at international conferences and workshops (see CV), ten invited talks at international universities (including Nancy, King's College, and Edinburgh), and, in a clear recognition from the international scientific community, four invited talks at international workshops in Spain, Israel, the US, and Japan.

Finally, I am actively involved in international research organization and evaluation activities. I have significant responsibilities within the Association for Computational Linguistics (ACL), the main research organization of my field: I currently serve as the editor of a Special Issue of the ACL journal *Computational Linguistics* (MIT Press); I have been in the SIGSEM Board of the ACL since 2013; I have acted as area co-chair for ACL 2016, program co-chair of *SEM 2015, and area co-chair of *SEM 2013 (*SEM is jointly organized by two Special Interest Groups of the ACL); and I regularly review for the top journals and conferences in the field (among others, ACL, EMNLP, COLING, LREC, EACL; *Computational Linguistics*, *Artificial Intelligence*, *Natural Language Engineering*, *Language Resources and Evaluation*, all of them indexed in the JCR). My responsibilities extend beyond the ACL: For instance, I am on the Editorial Board of the *Linguistic Issues in Language Technology* (LILT) journal (associated with the Linguistic Society of America) and I have evaluated projects for the national research agencies of Argentina and Poland.

5. Other merits (*Resto de méritos*).

Prizes and mentions. For my academic excellence, I obtained the *Premio Extraordinario de fin de carrera* (End of Degree Award) in Spanish Philology in 2000 and a *Mención especial* in the *Premios Nacionales de Fin de Carrera de Educación Universitaria* (Special Mention, National Prizes for End of Degree in University Education) of the Spanish Ministry. An award I was especially honored to receive is the Outstanding Reviewer Recognition at ACL 2015, since I take my reviewing work very seriously and I like to help other researchers make their work better and easier to understand.

Public engagement activities. I strive to impact society with my research. For this reason, I have carried out public engagement activities to disseminate Computational Linguistics to the general public and to promote the social use of computational tools and language resources, publishing three articles in general interest journals (refs. 47-49) and co-organizing an inter-sectorial event on the computational processing of Catalan with over 150 participants from research, industry, and administration: the *Jornada del Processament Computacional del Català* (Workshop on the Computational Processing of Catalan, Barcelona, March 2009).

Also because of this interest, I have actively pursued the construction of freely available computational tools and resources for Spanish and Catalan: To mention the two most prominent examples, I collaborated in the adaptation of the general FreeLing open source processing tool to Old Spanish and Old Catalan[5], and I led the construction of a 120-million word for Spanish and a 50-million word corpus for Catalan; they both can be freely downloaded and used.[6]

Teaching. Finally, I have ample teaching experience at the B. A. and Master's levels in Spain and abroad, both in Computational Linguistics (seven different courses on general Computational Linguistics, Computational Pragmatics and Semantics, Machine Translation, and related topics at the Universidad Pompeu Fabra) and Linguistics (two courses at the Universidad Pompeu Fabra and one at The University of Texas at Austin). I have also been actively involved in the European Summer School in Logic, Language and Information (ESSLLI), the top European linguistics summer school: I co-taught a course on Computational Lexical Semantics at ESSLLI 2009 (Bordeaux, France), I acted as local co-chair for ESSLLI 2015, and I am currently organizing a workshop at ESSLLI 2016. I have successfully combined my research activities with my teaching, and tried to transfer as much as possible of what I do to students.

6. Leading my own research line (*Capacidad del candidato para liderar su línea de investigación*)

During my career I have shown great leadership capabilities, as well as independence in my research, as follows. Already as a PhD student, I started having an **independent research line** with respect to my PhD supervisors, engaging in research on Computational Linguistics and semantic theory with colleagues at U. Pompeu Fabra and abroad (refs. 33, 34, 42). I have later continued to build a research line of my own at the intersection between Computational Linguistics and semantic theory, profitably using my independent working positions funded through competitive fellowships (see below). To pursue the kind of interdisciplinary research that characterizes me, I have actively sought collaborators that complement my knowledge and skills, in different fields (Linguistics, Computer Science, Cognitive Science, Physics) and geographical locations (Spain, Germany, Italy, US). As a result, I have co-authored almost half of my publications with researchers outside my institution. This also showcases my **excellent team work and team coordination capabilities**, also shown by the fact that I have been invited to participate in 17 research projects at the regional, national, European, and international levels. I have often taken coordinating roles in these projects: For instance, I was in charge of organizing the bi-weekly meetings of the US DARPA project I participated in at The University of Texas at Austin. Finally, I have successfully worked together with dozens of international researchers in the research, organization, and management activities mentioned in the previous sections.

As the above suggests, I am a **natural leader**: I am the Principal Investigator of a EU Marie Sklodowska-Curie grant; as detailed in Section 4, I serve in high-profile management roles in international scientific events and organizations in Computational Linguistics, including journal editing; I am the first author in almost 40% of my publications (19/49); moreover, in another 11 publications the first author was a student mentored by myself (refs. 3, 15, 17, 20, 41) or by some of my national and international collaborators (refs. 5, 7, 8, 10, 11, 16). Regarding student supervision, note that my post-doc positions have often hindered me from officially supervising students, either because of their short duration or because of university regulations. Despite this fact, I have mentored six students, including one ongoing PhD thesis and two masters' theses.

Finally, I have self-funded almost all of my career with **competitive grant programs**: An early Erasmus-Sócrates study fellowship that led to my first research experience at the University of Cologne, a fellowship for the Introduction to Research by the CSIC whereby I spent three months at an Artificial Intelligence research center right after finishing my Spanish Philology degree, two PhD fellowships (one by the Generalitat de Catalunya, one by the Fundación Caja Madrid, also including competitive travel funds for two research visits to Saarland University), funding from the PASCAL2 European Network of Excellence and the German SFB 732 project for a 5-month research visit at the University of Stuttgart, one three-year post-doctoral contract in the *Juan de la Cierva* program of the Spanish government, one three-year fellowship/contract in the *Beatriu de Pinós* program of the Catalan government, and the aforementioned EU Marie Sklodowska-Curie grant.

[5] http://nlp.lsi.upc.edu/freeling.

[6] http://www.cs.upc.edu/~nlp/wikicorpus. See the other resources I have developed at http://gboleda.utcompling.com/resources.

7. Conclusion

I carry out research that is substantially advancing theoretical and computational semantics and impacting other fields such as Physics. I have built a solid research career with ample international experience, and I am recognized as an established member of the Computational Linguistics community. My career is based on my broad research interests and theoretical and computational capabilities, on the one hand, and my strong leadership, team work, and coordinating skills, on the other. A *Ramón y Cajal* fellowship will provide me with the necessary framework and time frame to further develop my own research line and consolidate my leading position at the international level, and it will represent a key step on the path to a permanent position. In turn, I will significantly contribute to Spain's international recognition as a reference point for computational semantics and semantic theory, and to the European science space more generally: The fellowship will open up new research and funding opportunities with the collaborators in my research network encompassing Italy, Germany, and the US.

References

Alexiadou, A. and M. Stavrou. 2011. Ethnic Adjectives as pseudo-adjectives: a case study on syntax–morphology interaction and the structure of DP. *Studia Linguistica* 65(2):117-146.

Asher, Nicholas. 2011. *Lexical meaning in context: A web of words*. Cambridge University Press.

Baeza-Yates, R. and B. Ribeiro-Neto. 2011. *Modern Information Retrieval*. Addison-Wesley: Wokingham, UK. 2nd edition.

Baroni, M., R. Bernardi, R. Zamparelli. 2014. Frege in space: A program for compositional distributional semantics. *Linguistic Issues in Language Technologies* 9(6): 5-110.

Coecke, B., M. Sadrzadeh, S. Clark. 2011. Mathematical Foundations for a Compositional Distributional Model of Meaning. *Linguistic Analysis* 36: 345-384.

Erk, K. 2013. Towards a semantics for distributional representations. *IWCS 2013*, pp. 95-106, Potsdam, Germany.

Fábregas, A. 2007. The internal syntactic structure of relational adjectives. *Probus* 19(10):135–170.

Feng, Y., M. Lapata. 2010. Visual information in semantic representation. *NAACL 2010*, 91–99.

Harnad, S. 1990. The Symbol Grounding Problem. *Physica D* 42: 335-346.

Landauer, T. and S. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review* 104(2):211–240.

Mitchell, J., M. Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science* 34(8), 1388–1429.

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

Portner, P. 2005. *What is Meaning? Fundamentals of Formal Semantics.* Wiley-Blackwell.

Pustejovsky, James. 1995. *The generative lexicon*. The MIT Press.

Silberer, C., V. Ferrari, M. Lapata. 2013. Models of Semantic Representation with Visual Attributes. *ACL 2013*, 572-582.

Socher, R., B. Huval, Ch. Manning, A. Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. *EMNLP 2012*, 1201-1211.

Turney, P. and P. Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research* 37, 141–188.