Part B Template

START PAGE

MARIE SKLODOWSKA-CURIE ACTIONS

Individual Fellowships (IF)
Call: H2020-MSCA-IF-2014

PART B

"LOVe"

"**L**inking **O**bjects to **Ve**ctors in distributional semantics: A framework to anchor corpus-based meaning representations to the external world"

This proposal is to be evaluated as:

[Standard EF]

TABLE OF CONTENTS

List of Participants

Participants	Legal Entity Short Name	Academic (tick)	Non-academic (tick)	Country	Dept./ Division / Laboratory	Supervisor	Role of Partner Organisation
Beneficiary							
- University of Trento	UNITN	X		Italy	Center for Mind/Brain Sciences (CIMeC), Language Interaction and Computation Laboratory (CLIC)	Marco Baroni	

1. Summary

Language mediates between concepts in our mind and the things they refer to in the world. Semantic theories are typically biased towards conceptual or referential aspects. **My goal is to develop a theory of meaning that takes both aspects into account, and is supported by computational modelling experiments**, so that it will also enable computers to match linguistic expressions with entities in the world. This is a highly interdisciplinary proposal that **will bring computational linguistics, artificial intelligence, and theoretical linguistics forward.**

My model is based on distributional semantics, a scalable and flexible approach to computational semantics that, by inducing meaning representations from naturally occurring data with statistical methods, can model large portions of the lexicon and account for nuances in meaning that pose difficulties to traditional semantic theories. Distributional semantics has so far largely eschewed the reference issue, by testing its models on language-internal tasks. The project **bridges this language-world gap, and integrates the distributional framework into a referential semantic theory.** The project promises to advance our scientific understanding of language, a defining trait of the human species, and to make significant progress towards building computers we can talk to, with the ensuing strong impact on our everyday lives.

Even though I am an established researcher in computational semantics and also contributed to semantic theory, I still need to fully develop my own line of research to become a leading, independent researcher in Europe. Carrying out the present proposal at the University of Trento CLIC laboratory will be a fundamental step towards achieving my goal, since CLIC is a world leader in distributional semantics. Conversely, my unique profile, addressing theoretical linguistic questions through computational means, will fill a gap in the lab, widening the scope and outreach of the research conducted at CLIC.

2. Excellence

2.1 *Quality, innovative aspects and credibility of the research (including inter/multidisciplinary aspects)*

Introduction, state-of-the-art, objectives and overview of the action. Humans use language to refer to things in the world. However, language abstracts away from many contingent details to be able to describe a potentially infinite variation in reality (we apply the same word, *musician*, to Wolfgang Amadeus Mozart and to John Lennon), connecting our conceptual structure with the world. Semantic theories are typically biased towards referential or conceptual aspects of meaning. My goal is to develop a theory of meaning that takes both conceptual and referential aspects into account, and is supported by computational modelling experiments, so that it will also enable computers to match linguistic expressions with entities in the world. For instance, we will model the different meaning of *red* when used for a car or a nose (Figure 1).

Figure 1. ***Red car* vs. *red nose.***

Indeed, there has traditionally been a divide between formal semantics in the Montagovian tradition,[1] which focuses on truth and reference, and other approaches such as distributional semantics, which focus on more conceptual aspects of meaning. Formal semantics is very well *grounded* in external reality: For instance, *musician* refers to set of musicians there are in a given model of the world, say, John, Paul, George, and Ringo. However, formal semantics has very little to say about how this link between words and entities comes about in the first place: What properties does John have such that he qualifies as a musician, and how do words and phrases mirror these properties? As a consequence, formal semantics has trouble handling semantic phenomena that require

[1] Thomason, R. H. (Ed.). (1974). *Formal philosophy: Selected Papers of Richard Montague*. New Haven: Yale University Press.

access not only to the entities but also to their properties, such as similarity (*musician* is more similar to *singer* than to *professor*). Also for this reason, even if it comes with a well-defined *composition* mechanism to build the meaning of phrases (*a red car*) from the meaning of their component words (*a*, *red*, *car*), it struggles to account for the meaning modulations caused by modification, such as the fact that a red nose is of a very different shade than a red car (see Figure 1). Last but not least, this "direct" grounding of words also means that formal semantics does not afford a learning mechanism to induce semantic representations from data, and so it cannot scale up to realistic vocabulary sizes.

Distributional semantics,[2] an approach popular in computational linguistics and cognitive science that induces meaning representations from the contextual distribution of words in large amounts of linguistic data, exhibits complementary strengths and weaknesses: It provides a very rich and flexible semantic representation of words that is automatically constructed, so it can easily scale up and it excels at modelling word similarity. Moreover, recently, *compositional distributional semantics* has been the focus of a great research effort[3] –and in fact the host group of the present fellowship, the CLIC lab at CIMeC (University of Trento), is a world leader on the topic. This research has shown that distributional semantics naturally accounts for sense modulations of the *red car – red nose* type.[4] However, the link to the world remains a virtually unexplored challenge: Distributional semantics might tell you that a musician is in general similar to a singer, but it has currently no way to say whether a specific individual in the world qualifies as a singer. Consequently, it is typically applied to language-internal tasks only.

To sum up, **semantic theory is in need of a conceptually rich framework that links semantic representations to entities in the world in a principled manner**, and current theories only fragmentarily account for this link. Providing this framework is the ultimate goal of this project. A distinctive feature of my research is the fact that its goals are theoretical but its approach is computational, testing the theory with computer simulations and large-scale experiments. This makes it relevant to artificial intelligence and, ultimately, practical applications, in this case because the language-world link is crucial for systems that require reference to the external world, from geolocation to robotics. My research is thus genuinely interdisciplinary, impacting at this stage computational and theoretical linguistics and artificial intelligence, and at a future point also potentially computer-based engineering at large.

The project specifically aims at answering these three related questions:

- **Question 1**: Can distributional representations of noun phrases like *most influential band of the 60s* be linked to entities in the real world like The Beatles?
- **Question 2**: Can distributional representations of noun phrases containing visual attributes (*a red nose*) be linked to images described by these noun phrases?
- **Question 3**: Can formal and distributional semantics be combined into an overarching theory encompassing conceptual and referential aspects of word meaning?

I will tackle the first two questions through computational modelling experiments, and the last one by building on the empirical results of the experiments to propose a theoretical framework, as detailed next.

Research methodology and approach. Distributional semantics models word meaning as a function of the contexts in which words occur, by processing large amounts of data and recording how many times a given word occurs in a given context. A context for a word can be for instance a neighbouring word or a document it occurs in. A semantic representation for a word then essentially consists of a list of context counts, or *vector*. Any two words can then be compared using well defined linear algebra techniques, to induce, for instance, that

[2] Landauer, T., 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.; Turney, P., P. Pantel (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research* 37, 141–188.

[3] A.o.: Mitchell, J., M. Lapata (2010). Composition in distributional models of semantics. *Cognitive Science* 34(8), 1388–1429; Baroni, M., R. Bernardi, R. Zamparelli (2013). Frege in space: A program for compositional distributional semantics. *Linguistic Issues in Language Technologies* 9(6): 5-110.

[4] Boleda, G., M. Baroni, N. The Pham, L. McNally (2013). Intensionality was only alleged: On adjective-noun composition in distributional semantics. IWCS 2013, 35-46.

singer and *musician* are very similar in meaning, while *professor* and *cucumber* are not, just because humans use the former in similar contexts, the latter in different contexts. Furthermore, vectors for words can be combined, again through algebraic operations, to yield representations for phrases or even sentences (see fn. 3). I will carry out the following experiments to enable distributional semantics to handle referential aspects of meaning.

Question 1 will be addressed in **Experiment 1**, through an experiment mapping noun phrases like *most influential band of the 60s* to entities in a database (FreeBase), which contains rich ontological information. The entities will be represented: (1) As surrogates of real-world individuals, by vectors recording the information in FreeBase (e.g., Mozart has *Composer* as value for slot *Property*).[5] (2) Linguistically, by distributional vectors recording their use in a textual corpus, using a standard Named Entity Recogniser. We will next induce a mapping from one type of representation to the other, by relying on cutting-egde deep learming architectures of the sort that are currently being explored at CLIC. The mapping will be tested and optimized on a held-out development set prior to the main experiment.

The main experiment test set will consist of 400 entities evenly distributed across 10 relevant FreeBase categories, paired with two noun phrases of similar length and syntactic complexity that either uniquely describe the entity (*most influential band of the 60s*) or apply to several entities in a given FreeBase category (*four-member rock band from the UK*). We will attempt the ambitious task of connecting linguistic vectors representing the noun phrases to FreeBase entities by constructing noun phrase representations through advanced compositional distributional semantic methods (also developed at CLIC) and projecting them onto FreeBase space with the mapping function learned above. Our evaluation will test: (1) Are we getting the right entity (e.g., The Beatles) when given the representation for the noun phrase (*most influential band of the 60s*)? (2) Since we want reference, and not merely similarity, can we identify when the noun phrase picks a unique entity, as in the previous example, and when it applies to a set, as in *four-member rock band from the UK*? We will evaluate these questions through top-N precision and inducing a distance threshold between phrases and entities they might refer to, respectively.

Experiment 2 will address Question 2, linking noun phrases containing visual attributes (*red car*, *red nose*) with the right real-life pictures (see Figure 1), where the latter are represented through automated image analysis techniques that I have already experimented with.[6] The visual attributes will be expressed through three different kinds of nominal modifiers: Adjectives, nouns, and prepositional phrases. In this experiment, it is crucial that the compositional distributional methods capture how the meaning of the modifier (*red*) depends on the noun, that is, the visual correlate that can be expected to be present in the image. Previous research[7] has only tested highly transparent, compositional cases of the *red car* type; also, it has only tested whether the attribute is true, not whether it is salient: For instance, a human would not describe the first image in Figure 1 as *car with wheels*, because all cars have wheels. Therefore, a major contribution of this experiment will be the construction of the evaluation dataset itself, which will test linguistically interesting kinds of composition and will contain, in a controlled manner, false, true-and-salient, and true-but-non-salient attributes.

We will take 500 images from ImageNet,[8] controlling for the presence of both certain nouns (automatically) and certain modifiers of interest (manually). The images will be annotated in a crowdsourcing experiment through an online service[9] that puts scientists in touch with a large pool of volunteer participants for online surveys. We will gather two types of annotation: A natural description of the image, to gather salient modifiers (*red car*), and a true but non-salient description, like *car with wheels*. The false attributes will in turn be

[5] FreeBase: http://www.freebase.com. This type of world representation has precedents e.g. in Q. Cai, A. Yates (2013). Semantic Parsing Freebase: Towards Open-domain Semantic Parsing. **SEM 2013*.

[6] Bruni, E., G. Boleda, M. Baroni, N. K. Tran (2012). Distributional semantics in technicolor. ACL 2012, 136-145.

[7] Russakovsky, O., L. Fei-Fei (2012). Attribute Learning in Large-scale Datasets. ECCV 2012 Workshop.

[8] http://www.image-net.org, freely available and widely used image database aligned with WordNet.

[9] http://crowdflower.com. See Section 7 (ethical aspects) for more information.

obtained by randomly permuting across images and manually checking for falsity. The model and the evaluation will be similar to the one in Experiment 1, with two new challenges: The use of visual information from real-life pictures instead of the more controlled database information, and the identification of salient features.

Experiments 1 and 2 will yield viable mechanisms to bridge the language-world gap in distributional semantics, as well as a wealth of data, allowing us to address Question 3: Their results will inform the development of a **semantic framework** with a mutually beneficial division of labour between distributional semantics (conceptual aspects) and formal semantics (referential aspects). Previous research combining the two approaches[10] has used distributional models to add weighted rules or create macro-predicates within a logical framework. We will achieve a fuller integration with a larger coverage of empirical phenomena by relying on distributional semantics to link noun phrase vectors with entities, as in Experiments 1 and 2, and on formal semantics to deal with the ensuing referents in a symbolic fashion. For instance, in the following dialogue:

Adam *(pointing to an image in a magazine)*: John has bought that red car.

Barbara: It must have been expensive!

the specifications for how the red car is to be identified are provided by distributional semantics, and the anaphoric *it* is handled via standard formal semantics mechanisms (e.g., in Discourse Representation Theory[11]).

A large component of this part of the project is dissemination, reaching out to the theoretical linguistics community to illustrate more clearly how distributional semantics is relevant to semantic theory: First, how it can address known challenges in semantics, second, the specific framework developed in the present project (see Section 3.2).

Originality and innovative aspects of the research programme. The strongest point of the ambitious and interdisciplinary research programme just presented is that it is original and innovative from both a computational and a theoretical standpoint. On the computational side, it extends cutting-edge tools in distributional semantics (compositional methods, cross-representational mapping, deep learning) to a new topic, reference. Previous research[12] tackling computational approaches to reference has mainly been in the subfields of natural language generation, semantic parsing, and computer vision. This research involves either very limited and artificial virtual worlds or a very limited linguistic component, and does not address phenomena such as meaning modulation. Because of the large-scale vocabulary coverage and successful composition mechanisms distributional semantics affords, and given recent advances in image analysis, it is now time to tackle reference in a less constrained fashion. My proposal goes beyond the state of the art in handling a sophisticated representation of the world through databases, images, and text, and tackling a wider range of natural language expressions.

I have experience in both distributional and theoretical semantics. This research proposal will place me at the forefront of research on the computational-theoretical semantics interface, because it addresses significant gaps in formal and distributional semantics in an integrated way that will advance both fields by building on their complementary strengths. The publications ensuing from the project and the experience at the world leading CLIC lab will be crucial for my career, enabling me to reach a tenured or tenure-track position in Europe after the fellowship (see Section 2.2). In turn, this project will open up new collaboration opportunities for the CLIC lab, thanks to the research network I have built in the past, encompassing Germany, Spain, and the US (see Section 3.1), with stable collaborations on topics related to the present project.

[10] Beltagy, I., C.K. Cuong, G. Boleda, D. Garrette, K. Erk, R. Mooney (2013). Montague meets Markov: Deep semantics with probabilistic logical form. *SEM 2013;* Lewis, M., M. Steedman (2013). Combined Distributional and Logical Semantics. *TACL* 1:179–192.

[11] Kamp, H., U. Reyle (1993). *From Discourse to Logic.* Kluwer: Dordrecht.

[12] E.g., Chen, D.L., J. Kim, R. Mooney (2010). Training a Multilingual Sportscaster: Using Perceptual Context to Learn Language. *Journal of Artificial Intelligence Research* 37:397-435; Krahmer, E., K. van Deemter (2012). Computational Generation of Referring Expressions: A Survey. *Computational Linguistics* 38(1):173-218.

2.2 Clarity and quality of transfer of knowledge/training for the development of the researcher in light of the research objectives

The *Experienced Researcher* will gain new knowledge from the hosting organisation through training. My overall scientific training objective is **(TO1) to acquire the necessary skills to become an expert working at the theoretical-computational semantics interface.** I acquired a solid humanist background during my B.A., and skills during my PhD and post-doctoral experience that allowed me to significantly contribute to computational linguistics and semantic theory (see CV). To carry out the ambitious and interdisciplinary research programme just outlined, I need to deepen my knowledge of advanced algebra and statistics, acquire further hands-on programming training, and gain more expertise in formal semantics. My scientific training subgoals involve learning how to:

- TO1.1, implement and use deep neural network representation learning and cross-representational mapping techniques;
- TO1.2, master recent advances in compositional distributional semantics;
- TO1.3, process visual (image) data;
- TO1.4, process large-scale data with cluster-based and GPU computing;
- TO1.5, build semantic datasets through crowdsourcing;
- TO1.6, extend my command of formal semantic tools for linguistic analysis.

The CLIC lab is the ideal place to solidify these skills, since it is a leader in compositional distributional semantics (largely thanks to Prof. Baroni's ERC Grant COMPOSES), including the creation of datasets; it is currently porting distributional semantics to deep learning; it has ample expertise in combining visual and textual distributional data; and two of its tenured researchers (R. Zamparelli and R. Bernardi) are widely recognized theoretical linguists. Daily hands-on training will be ensured by project-related collaborations, individual meetings, collaborative tutorials, and technical seminar sessions with highly qualified post-doctoral (G.Dinu, D. Paperno) and pre-doctoral researchers (A. Lazaridou, G. Kruszewski, N. Pham).

My strategic training objective is **(TO2) to become a leading, independent researcher**, ready to attain a stable position in Europe and apply for funding as a PI, learning how to:

- TO2.1, produce higher impact publications and to publish more in journals (computational linguistics is conference-oriented, while scientist evaluations are often journal-based);[13]
- TO2.2, publish outside my main field of expertise (computational linguistics), reaching out to artificial intelligence and linguistic audiences;
- TO2.3, expand my mentoring experience;
- TO2.4, enhance my public engagement skills;
- TO2.5, acquire hands-on project management training in aspects like budget or ethics.

CLIC ensures my training in these skills because its researchers engage in interdisciplinary research and routinely publish in high-impact journals in and outside my field (e.g. *Computational Linguistics*, *Linguistic Inquiry*, *Journal of Artificial Intelligence Research*, *Journal of Cognitive Neuroscience*); it offers mentoring possibilities to its post-docs (I plan to supervise a master's thesis related to Experiment 2); it facilitates public engagement activities both through general programs at the University of Trento and through the support of the institute's dedicated press media agent (see Section 3.2); and it provides excellent support for grant management (see Section 4.2). Indeed, the fellowship will provide for great project management training, especially regarding (a) budget, because it includes funds in addition to the salary costs, and (b) ethics, as I will need to deal with procedures regarding ethics for the first time.

The *Experienced Researcher* has the capacity for transferring the knowledge previously acquired to the host organisation. What makes my profile unique is that, while I have a solid training in linguistics and my research goal is to find out more about how language works, I also have a keen interest and good training in computational linguistics, which allows me to address linguistic research questions through computer simulations and test them in large-scale experiments with a sound evaluation methodology. I will transfer to CLIC my

[13] See http://scholar.google.es/citations?view_op=top_venues&hl=en&vq=eng_computationallinguistics for the field's metrics.

knowledge and know-how about how to design and interpret computational experiments that pursue genuinely linguistic questions.

2.3 Quality of the supervision and the hosting arrangements

Qualifications and experience of the supervisor. **Marco Baroni** carries out cutting-edge, interdisciplinary research at the interface between computational linguistics, cognitive science, and artificial intelligence. 27 of his over 100 publications appeared in high-impact international journals (including *Computational Linguistics*, *Cognitive Science,* and *Journal of Artificial Intelligence Research*) and ten at ACL, the most prestigious conference in computational linguistics. He has been the PI of two Italian national projects, a Google Research Award in 2010 on connecting distributional semantics and visual data, and an ERC Starting Grant in 2011 on compositional distributional semantics. The two latter grants are closely related to the present proposal, as is the ICT COST Action on Integrating Vision and Language of which he is a national representative and dissemination officer. He also has ample mentoring experience, including eight PhD students and four post-doctoral researchers. Prof. Baroni is widely recognised as a leader in computational semantics by the international community: A.o., he has given seven invited keynote talks (e.g. at KONVENS 2012), has been asked to contribute two surveys to *Language and Linguistics Compass*, and is in the Editorial Board of *Computational Linguistics*, the leading journal in the field.

Integration of the Experienced Researcher. I started collaborating with the CLIC group in 2010, through a project at U. Pompeu Fabra, and then with research related to the COMPOSES ERC Grant, which resulted in three high-impact conference proceeding papers, the co-organisation of a scientific workshop, and two workshop presentations (see CV). This previous collaboration ensures a smooth integration in CLIC, such that I will be able to engage in research from day one. Further dialogue, meeting, training, and networking opportunities will be provided by the regular scientific events at CLIC: The weekly COMPOSES meetings and reading group, the weekly CLIC series for internal discussion and presentation dry-runs, the bi-weekly CLIC Colloquium with talks by external researchers, and on-demand collaborative tutorials. Practical arrangements will be overseen by the CIMeC administration, with an extensive experience in hosting foreign researchers (also see Sections 4.3 and 6): I will have access to an office, all the necessary equipment for the completion of the project, IT and administrative support, and support for housing (university lodging for the first three months) as well as meal coupons. Throughout the fellowship, I will be fully covered by the Italian legislation, providing standard social security benefits.

A detailed Career Development Plan will be designed together with the supervisor prior to the start of the fellowship, and monitored throughout, to ensure the achievement of the short- and long-term objectives described in Section 2.2. It will cover, a.o.:

- Research results: Publication and conferences / workshops as in Section 3.2.
- Research skills and techniques: Training objectives TO1.1-TO1.6 in Section 2.2.
- Research management: TO2.5, plus hands-on training by applying for further funding during the fellowship (travel awards, etc.) and after (see Section 3.1).
- Communication skills: TO2.1, TO2.2, and TO2.4.
- Other professional training: ACL and related conferences tutorials, UNITN crash course on intellectual property and knowledge transfer, UNITN workshop on external funding.
- Anticipated networking opportunities: Conferences, workshops (see Section 3.2), and the world-renowned researchers that regularly visit CLIC and CIMeC.

2.4 Capacity of the researcher to reach and re-enforce a position of professional maturity in research

During my career I have shown great potential for reaching a position of professional maturity, as follows (see CV). I have **self-funded** almost all of my career with competitive grant programmes including two PhD and two post-doctoral fellowships. I am a **natural leader** and I have **excellent team work and team coordination capabilities**: I have been elected as a member of the Department Council both as an undergraduate and as a

graduate student; already as a post-doc, I have mentored four students, including two masters' theses; I have co-organised two international scientific events; I have worked with a broad variety of collaborators, co-authoring almost half of my publications with researchers outside my institution; and I have participated in 14 research projects at the regional, national, European, and international levels. My research has a notable **impact** (most cited publication: 101 citations, h-index: 11).[14] As a result, I am **recognised by the international community**: I will be in the Editorial Board of the *Linguistic Issues in Language Technologies* journal as of fall 2014, I have started to manage scientific events and organisations (area co-chair for *SEM 2013, Information Officer in the SIGSEM-ACL board, local co-chair for ESSLLI 2015), and I have given ten invited talks at international universities, including Nancy (France), King's College (London, UK), and Saarland (Germany). The time is ripe for my career to blossom, both in terms of the impact of my research and the professional position I will be able to reach in the next step of my career.

3. Impact

3.1 *Enhancing research- and innovation-related human resources, skills, and working conditions to realise the potential of individuals and to provide new career perspectives*

This fellowship will enable me to become a **leading, independent researcher in Europe**: In the short-term, I will acquire crucial competencies and develop an autonomous line of research (see Section 2.2); in the mid- and long-term, I expect to obtain a tenured or tenure-track position and apply for funding as a PI. Concretely, the next step in my career will be to apply for an ERC Grant, for tenure-track programmes such as the Spanish *Ramón y Cajal*, and for university lecturer or professor positions. The fellowship will represent a huge leap in my career, because it will make my profile very competitive by enabling me to publish in high-impact venues and gain more visibility (see Section 3.2), as well as enriching my research network with an Italian dimension. In my previous PhD and post-doctoral experience I established a solid network in Spain, Germany, and the US. My main collaborators in these countries, L. McNally, S. Padó, S. Schulte im Walde, and K. Erk, share research goals with the CLIC lab, and will be able to engage in common projects and apply for funding during and after the fellowship. This will be beneficial for the CLIC lab and for the European science base more generally, and potentially in later stages also for the European economy, given the connections of my research to practical applications: Language is the most natural communication means for humans, and this project makes progress in allowing people to talk to computers, thus (1) addressing the digital divide, (2) making daily operations such as using a GPS easier for everybody.

3.2 *Effectiveness of the proposed measures for* ***communication and results dissemination***

Communication and public engagement strategy of the action. My communication strategy is based on two parallel lines of dissemination: To the scientific community (see next paragraph), and to society at large through public engagement activities. As for the latter, in the past I have carried out public engagement activities (see CV), for instance participating in four press media publications and co-organising an inter-sectorial event on the computational processing of Catalan with over 150 participants from research, industry, and administration. I will expand on this experience during the fellowship, profiting from the activities organized by UNITN and the support of CIMeC's dedicated press agent, with presentations in:

- General Italian media (e.g. the weekly scientific supplement *Tuttoscienze* of *La Stampa*, Radio 3 Scienza, the *Superquark* show on the Italian television RAI 1);
- International media (e.g. one of the many that have previously reported CIMeC research);
- European Researchers' Night, with an interactive demo about Experiments 1 and/or 2.

[14] According to Google Scholar, see http://scholar.google.com/citations?user=NFJ9kUEAAAAJ&hl=en.

- Social networks, through the CIMeC FaceBook and UNITN LinkedIn pages;[15]
- ESSLLI 2016, with a course proposal on distributional semantics and linguistics.

Dissemination of the research results. As the project is highly interdisciplinary, the dissemination of its results to the scientific community will also be interdisciplinary, targeting:

- Computational linguistics: one high-impact conference and one journal article (e.g. ACL, EACL, EMNLP and TACL, Computational Linguistics) on Experiment 1;
- Artificial intelligence: one high-impact conference article (e.g. NIPS, IJCAI) on Experiment 2, which encompasses computer vision, and one overarching journal article integrating the results of both experiments (*Journal of Artificial Intelligence Research*);
- Theoretical linguistics: two journal articles (e.g., *Language*, *Lingua*, *Linguistics and Philosophy*), one at the beginning of the fellowship, on distributional semantics as a tool for linguistic research, one towards the end, on the framework developed in this project; at least one high-profile semantic conference (SALT, *Sinn und Bedeutung*); and three theoretical linguistics workshops (to be identified).

I will also present my research through invited talks at universities or scientific events whenever the possibility arises. The planned dissemination will very positively impact my career, as I will actively network and become highly visible at conferences and workshops, which in turn will foster a higher impact of the journal publications. Because it is an ambitious but well planned project (see Section 2.1) relevant to three neighbouring fields, it will advance the state of the art in each of these fields, and it will especially foster methodological innovation in theoretical linguistics.

Exploitation of results and intellectual property. Since my project consists of basic research, I do not expect to exploit the results commercially at this point. The only intellectual property issue I will need to handle concerns the dataset created in WP2, which I will make freely available for research purposes. I will have the support of the UNITN dedicated Knowledge Transfer Office to handle this and any other IPR issue that may arise. I am aware that under Horizon 2020 I will need to ensure open access to the publications arising from the present project, and I will implement the necessary measures: (1) provide immediate open access upon publication if possible, using project funds if necessary; (2) using the UNITN open access repository (UNITN e-prints) whenever appropriate.

4. Implementation

4.1 ***Overall coherence and effectiveness of the work plan,*** *including appropriateness of the allocation of tasks and resources*

Table 1 summarizes the work plan; the reminder of the section provides details.

Month	4	8	12	16	20	24
Work package 1						
Work package 2						
Work package 3						
Deliverable		D1.1	D2.1	D2.2		D3.1, D3.2
Milestone	M1	M2	M3	M4		M5
Conference		SALT…	ACL…		NIPS…	
Workshop		W1		W2		W3
Dissemination	S3.1, S3.2	S1.1, S1.2		S2.1, S2.2		S3.2
Public engagement			Italian media	ESSLLI	NIGHT	International media

Table 1. Gantt chart of the work plan.

[15] https://www.facebook.com/cimec.unitn, https://www.linkedin.com/company/university-of-trento.

Work Packages description. Each of the research questions described in Section 2.1 will be addressed in a dedicated Work Package, as follows (including timeline):

WP1 (Experiment 1): Identifying entities		
MONTHS	SCIENTIFIC GOAL	TRAINING GOALS
01 - 08	To gauge the referential potential of distributional representations by linking noun phrase vectors (*most influential band of the 60s*) to entities (The Beatles).	TO1.1 (deep learning), TO1.2 (compositional distributional semantics), TO1.3 (large-scale data), TO2.1 (high impact / journal publications), TO2.5 (budget).
WP2 (Experiment 2): Describing images		
MONTHS	SCIENTIFIC GOAL	TRAINING GOAL
09 - 16	To test the referential potential of distributional representations with a more challenging representation of the real world, linking noun phrase vectors containing visual attributes to images.	TO1.4 (image data), TO1.5 (crowdsourcing), TO2.1 (high impact / journal publications), TO2.3 (publish in artificial intelligence), TO2.3 (mentoring), TO2.4 (public engagement), TO2.5 (budget, ethics).
WP3: A conceptual and referential semantic framework		
MONTHS	SCIENTIFIC GOAL	TRAINING GOAL
1-4	To illustrate how distributional semantics is relevant to semantic theory.	TO2.1 (high impact / journal publications), TO2.3 (publish in linguistics).
17 - 24	To develop a semantic framework encompassing conceptual and referential aspects of meaning.	TO1.6 (formal semantics), TO2.1 (higher impact / journal publications), TO2.3 (publish in linguistics); TO2.4 (public engagement).

Table 2. Timeline of the work packages.

List of major deliverables. *D1.1*: WP1 report. *D2.1*: WP2 dataset (data deliverable). *D2.2*: WP2 report. *D3.1*: WP3 report. *D3.2*: project overview.

List of major milestones. *M1*: beginning of the fellowship. *M2*: end of WP1. *M3*: dataset completed. *M4*: end of WP2. *M5*: end of fellowship.

Conferences, workshops. See Section 3.2.

Dissemination. Timeline follows plan in Section 3.2. *S3.1*, *S3.2*: first journal and conference submissions for WP3. *S1.1*, *S1.2*: Conference and journal submissions for WP1. *S2.1*, *S2.2*: Conference and journal submissions for WP2. *S3.2*: second journal submission for WP3.

Public engagement. See Section 3.2. The timeline is approximate, pending event and fellowship dates; social media activities will be continuous and so are not included in Table 1.

4.2 Appropriateness of the management structure and procedures, including quality management and risk management

Project organisation and management structure. I will lead or take active part in all the management aspects of the project. I will meet Prof. Baroni bi-weekly to receive guidance regarding the training objectives, the scientific progress of the project, and risks that may endanger it (see next paragraph). I will also meet as appropriate with CLIC researchers I collaborate with, and during WP2 I plan to meet weekly with a master's student that will help out with WP2 tasks as part of his master's thesis. An External Advisory Board, including world-renowned experts Stephen Clark and Hinrich Schütze, will further monitor the quality of my research (including a visit to CLIC between milestones M2 and M4, funded by the fellowship). For the administrative management, I will be assisted by the expert UNITN Office of Scientific Research and Technological Transfer Division, which has an office at CIMeC that has already very efficiently assisted me in the proposal preparation. The CIMeC accounting department will assist me with financial management. I will meet with Office and accounting department representatives at every milestone (see Table 1), and whenever appropriate, to oversee the management of the project.

Risks that might endanger reaching project objectives.

RISK	CONTINGENCY PLAN
Problems in my integration in the CLIC lab (e.g., office, collaboration with colleagues).	Talk to Prof. Baroni to ask for support to solve them; implement the solutions.
Computational challenges in WP1 or WP2 (e.g. deep learning, image processing).	Allocate advising hours with CLIC member or computer vision specialist (Computer Science dept).
Crowdsourcing not working properly for WP2 research goals.	Redesign; if that fails, produce a smaller dataset with more controlled manual annotation.
Delays in the progress of the overall project due to problems in WP1 or WP2.	Give priority to the completion of WP2 so I can devote enough time to WP3.
Difficulties in dissemination of WP3 due to the linguistic community viewing distributional semantics as extraneous to its research goals.	Work together with Prof. Baroni, R. Zamparelli, and R. Bernardi to identify further adequate theoretical linguistics forums and revise manuscripts accordingly.

Table 3. Risks and contingency plans.

4.3 *Appropriateness of the institutional environment* (infrastructure)

The CIMeC (UNITN) CLIC lab numbers four faculty members, seven post-docs, and five PhD students. CLIC is one of the most active labs specialising in computational linguistics, emphasising links with theoretical linguistics/semantics and cognitive neuroscience. CLIC is part of a wider network of centres focusing on human language, knowledge and related areas in the Trento region. Besides collaborating with other groups at CIMeC, CLIC has close research and teaching ties with the UNITN Computer Science (featuring strong computer vision), Psychology and Humanities departments, the HLT group at the Trento FBK research institute, and the Trento-based Laboratory for Applied Ontology of the Italian CNR. In the past, CLIC has hosted one CIP European Project, 6 nationally- or regionally-funded projects, and one project sponsored by a Google Research Award. It is currently the host of an ERC Starting and an FP7 Cooperation Grant, and it is in the consortium of the Erasmus Mundus Master's in Language and Communication Technologies. Given its placement in CIMeC, its excellence in both theoretical and computational linguistics, and its focus on compositional semantics and on connecting language and vision, CLIC affords the maximum chance of a successful project outcome. The infrastructure needed for its completion (mainly cluster-based computing) is available at CLIC; see Sections 2.3 and 6 for more information.

4.4 *Competences, experience and complementarity* of the participating organisations and *institutional commitment*

The fellowship will be beneficial for both the Experienced Researcher and host organisation. (See sections 2.2 and 2.4 for further details.) The fellowship will be crucial to attain my training objectives: (TO1) to acquire skills to become an expert working at the theoretical-computational semantics interface, (TO2) to become a leading, independent European researcher. It will also add an Italian node to my international network (see Section 3.1). Conversely, CLIC will benefit from my network and unique profile; indeed, Prof. Baroni has published mainly in computational linguistics and cognitive science, tenured members R. Zamparelli and R. Bernardi are clearly on the linguistic side, and post-docs and PhD students more on the computational side. Because I address theoretical linguistic questions through computational means, my profile fills one gap that is central to the current research interests of CLIC. I will use my leadership and team work capabilities to promote tighter collaborations between theoretical and computational linguists. Moreover, the novel topic of reference within distributional semantics is complementary to current research in COMPOSES.

Commitment of beneficiary organisation to the programme. (Also see Section 6.) Given the fit between my profile and its current research goals, the CLIC and its head, Prof. Baroni, are fully committed to the success of the present proposal. Because of the added value of a Marie Curie fellowship, scientifically and financially, CIMeC is also committed to it, and will provide support in every aspect needed, from scientific dialogue to practical arrangements.

5. CV of the Experienced Researcher

PERSONAL INFORMATION

Last name, first name: Boleda, Gemma
Date of birth: 29/03/1976
URL for web site: http://gboleda.utcompling.com

• EDUCATION

2003 – 2007 PhD Universitat Pompeu Fabra
Universitat Pompeu Fabra, Spain
2000 – 2003 Master Cognitive Science and Language
Universitat Pompeu Fabra, Spain

• CURRENT POSITION

06/2014 – 05/2015 Post-doctoral researcher and lecturer
Department of Translation and Language Sciences, U. Pompeu Fabra, Spain

• PREVIOUS POSITIONS

04/2012 – 05/2014 Post-doctoral researcher and lecturer
Linguistics Department, University of Texas at Austin, USA
06/2011 – 03/2012 Researcher
Dept. of Translation and Language Sciences, U. Pompeu Fabra, Spain
04/2010 – 08/2010 Visiting researcher
Institute for Natural Language Processing, U. Stuttgart, Germany
06/2008 – 05/2011 Post-doctoral researcher
Dept. of Computer Science, U. Politècnica de Catalunya, Spain
05/2005 – 12/2007 Post-doctoral researcher
Barcelona Media Centre d'Innovació, Spain.
11/2004 – 12/2004 Visiting researcher
CoLi, Universität des Saarlandes, Germany
04/2003 – 06/2003 Visiting researcher
CoLi, Universität des Saarlandes, Germany
01/2001 – 12/2006 Doctoral researcher
Dept. of Translation and Philology, U. Pompeu Fabra, Spain
09/2000 – 12/2000 Research assistant
Artificial Intelligence Research Institute (CSIC), Spain
01/2000 – 06/2000 Student assistant
Dept. of Translation and Philology, U. Pompeu Fabra, Spain
11/1997 – 05/1998 Student assistant
Linguistic Information Processing group, Universität zu Köln, Germany

• FELLOWSHIPS AND AWARDS

2012 – 2015 *Beatriu de Pinós* post-doctoral fellowship, AGAUR, Spain
2008 – 2011 *Juan de la Cierva* post-doctoral fellowship, MICINN, Spain
2010 – 2010 PASCAL2 European Network of Excellence, Internal Visiting Programme, EU (funding for post-doctoral visit, U. Stuttgart, Germany)
2005 – 2006 PhD fellowship, *Fundación Caja Madrid*, Spain
2003, 2004 Short Research Visit Programme, AGAUR, Spain (funding for doctoral visit, Saarland U., Germany)
2001 – 2004 PhD fellowship, Catalan government, Spain
2001 Extraordinary Degree Award, U. Autònoma de Barcelona, Spain
2001 Honorable Mention, National Bachelor Degree Awards, Spanish Government
2000 Fellowship for the Introduction to Research, CSIC, Spain
1997 – 1998 Sócrates-Erasmus scholarship, EU

• SELECTED PUBLICATIONS

(Note: for space reasons, conference and workshop talks are not included in the CV)

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. Cit.: 14.1

Refereed journal articles (total: 9)

F. Font-Clos, G. **Boleda**, A. Corral. 2013. A scaling law beyond Zipf's law and its relation to Heaps' law. *New Journal of Physics* 15:9, 093033. Cit.: 9.

Boleda, G., S. Schulte im Walde, T. Badia. 2012. Modeling regular polysemy: A study in the semantic classification of Catalan adjectives. *Computational Linguistics* 38(3): 575-616. Cit.: 7.

Boleda, G., S. Schulte im Walde, T. Badia. 2008. An Analysis of Human Judgements on Semantic Classification of Catalan Adjectives. *Research on Language and Computatio*n 6(3): 247-271. Cit.: 2.

Mayol, L., G. **Boleda**, T. Badia. 2005. Automatic acquisition of syntactic verb classes with basic resources. *Language Resources and Evaluatio*n 39(4): 295-312. Cit.: 6.

Book chapters (total: 3)

Boleda, G., S. Evert, B. Gehrke, L. McNally. 2012. Adjectives as saturators vs. modifiers: Statistical evidence. In Aloni, M. et al. Eds*.): Logic, Language and Meaning - 18th Amsterdam Colloquium, Amsterdam, The Netherlands, December 19-21, 2011, Revised Selected Papers*. Lecture Notes in Computer Science 7218, pp. 112-121. Springer.

McNally, L., G. **Boleda**. 2004. Relational adjectives as properties of kinds. In Bonami, O. and P. Cabredo Hofherr (eds.) *Empirical Issues in Syntax and Semantics* 5, 179-196. **Cit.: 101**.

Articles in refereed conference proceedings (total: 16)

Roller, S., K. Erk, G. **Boleda**. 2014. Inclusive yet Selective: Supervised Distributional Hypernymy Detection. *CoLing 201*4, Dublin, Ireland.

Arsenijevic, B., B. Gehrke, G. **Boleda**, L. McNally. 2014. Ethnic adjectives are proper adjectives. In R. Baglini et al. (eds.), *CLS 46-I The Main Session: Proc. of 46th Annual Meeting of the Chicago Linguistic Societ*y, Chicago, IL, USA. Cit.: 5.

Beltagy, I., C. K. Cuong, G. **Boleda**, D. Garrette, K. Erk, and R. Mooney. 2013. Montague meets Markov: Deep semantics with probabilistic logical form. **SEM 201*3. Atlanta, US. Cit.: 12.

Boleda, G., M. Baroni, N. The Pham, L. McNally. 2013. Intensionality was only alleged: On adjective-noun composition in distributional semantics. *IWCS 201*3, Potsdam, Germany. Cit.: 5.

Boleda, G., E. M. Vecchi, M. Cornudella, L. McNally. 2012. First-order vs. higher-order modification in distributional semantics. *EMNLP-CoNLL 201*2, Jeju Island, Korea. Cit.: 8.

Bruni, E., G. **Boleda**, M. Baroni, N. K. Tran. 2012. Distributional semantics in technicolor. *ACL 201*2, Jeju Island, Korea. **Cit.: 31.**

Boleda, G., S. Padó, J. Utt. 2012. Regular polysemy: a distributional model*. *SEM 2012,* Montréal, Canada. Cit.: 9.

Sánchez-Marco, C., G. **Boleda**, J.M. Fontana, J. Domingo. 2010. Annotation and representation of a diachronic corpus of Spanish*. LREC 201*0, Valletta, Malta. Cit.: 14.

Boleda, G., S. Schulte im Walde, T. Badia. 2007. Modelling Polysemy in Adjective Classes by Multi-Label Classification. *EMNLP-CoNLL 200*7. Cit.: 10.

Boleda, G., T. Badia, E. Batlle. 2004. Acquisition of Semantic Classes for Adjectives from Distributional Evidence. *CoLing 200*4, Geneva, Switzerland. Cit.: 11.

Padó, S. and G. **Boleda**. 2004. The Influence of Argument Structure on Semantic Role Assignment. *EMNLP 200*4, Barcelona, Spain. Cit.: 5.

Edited volumes

Herbelot, A. and Zamparelli, R. and **Boleda**, G. (eds.). 2013. *Proc. of the IWCS 2013 Workshop Towards a Formal Distributional Semantic*s. Potsdam, Germany: ACL.

Artstein, R., G. **Boleda**, F. Keller, S. Schulte im Walde (eds). 2008. *Proc. of the COLING Workshop on Human Judgements in Computational Linguistic*s. Manchester, UK: CoLing 2008 Organizing Committee.

Articles in refereed workshop proceedings (total: 9)

Beltagy, I., S. Roller, G. **Boleda**, K. Erk, R. Mooney. 2014. UTexas: Natural Language Semantics using Distributional Semantics and Probabilistic Logic. To appear in *SemEval 201*4, Dublin, Ireland.

Sánchez-Marco, C., G. **Boleda**, L. Padró. 2011. Extending the tool, or how to annotate historical language varieties. *ACL 2011 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2011)*, Portland, Oregon, USA. Cit.: 14.

Boleda, G., S. Bott, C. Castillo, R. Meza, T. Badia, V. López. 2006. CUCWeb: a Catalan corpus built from the Web. *Second Workshop on the Web as a Corpus at EACL'0*6. Trento, Italy. Cit.: 11.

Badia, T., G. **Boleda**, M. Melero, A. Oliver. 2005. An n-gram approach to exploiting a monolingual corpus for Machine Translation. *Second Workshop on Example-based Machine Translatio*n, *MT Summit* X, Phuket, Thailand. Cit.: 20.

Boleda, G., T. Badia, S. Schulte im Walde. 2005. Morphology vs. Syntax in Adjective Class Acquisition*. ACL-SIGLEX 2005 Workshop on Deep Lexical Acquisitio*n, Ann Arbor, USA. Cit.: 9.

Boleda, G. and L. Alonso. 2003. Clustering Adjectives for Class Acquisition*. EACL 2003 Student Research Worksho*p, Budapest, Hungary. Cit.: 7.

Articles in press media (public engagement)

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

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

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

Badia, T., G. **Boleda**, M. Quixal. 2001. Curso sobre Tecnologías de la lengua (segunda edición). *QUARK, Ciencia, Medicina, Comunicación y Cultur*a 21, 14-16.

• INVITED TALKS

17/07/13 Intensionality was only alleged: On adjective-noun composition in distributional semantics. *Guest lecture series Sonderforschungsbereich 732: Incremental specification in contex*t, Stuttgart University, Germany.

06/09/11 Modeling regular polysemy: A study in the semantic classification of Catalan adjectives. *Linguistics Colloquiu*m, University of Texas at Austin, Austin, USA.

14/07/11 Modeling regular polysemy: A study in the semantic classification of Catalan adjectives. *CLIC Research Seminar* , CIMeC, Rovereto, Italy.

06/02/10 Word Sense Disambiguation and regular polysemy. *Institutsversammlun*g, Institut für Maschinelle Sprachverarbeitung, Stuttgart, Germany.

03/19/10 Computational Feedback to Linguistics: A study in the semantic classification of Catalan adjectives. *Nancy NLP Semina*r, INRIA-Lorraine, France.

11/14/07 Automatic acquisition of semantic classes for adjectives. *Natural Language Processing Semina*r, Universitat Politècnica de Catalunya, Barcelona, Spain.

06/28/05 Adquisició de classes semàntiques adjectivals. *III Workshop of the PhD Program in Cognitive Science and Language: "Acquisition*", Barcelona, Spain.

10/12/04 Acquiring Semantics Classes for Adjectives through Clustering. *Computational Linguistics Semina*r, King's College, London, UK.
11/25/04 A Quantitative Approach to the Lexical Semantics of Adjectives. *Computational Linguistics Colloquiu*m, Universität des Saarlandes, Saarbrücken, Germany.
11/11/04 Acquisition of Semantic Classes for Adjectives. *Colloquium of the International Post-Graduate College on Language Technology and Cognitive System*s, Universität des Saarlandes, Saarbrücken, Germany.

- **PARTICIPATION IN FUNDED PROJECTS (as collaborating researcher)**

2012 – 2014 Defense Advanced Research Projects Agency (DARPA), US, Deep Exploration and Filtering of Text (DEFT) Program, "Statistical Relational Learning and Script Induction for Textual Inference", $1,302,318. PI: Ray Mooney, Co-PI: Katrin Erk, University of Texas at Austin.
2011 – 2013 Spanish government, FFI2010-15006/FILO ("OntoSem2: Natural language ontology and the semantic representation of abstract objects 2"), €104,000. PI: Louise McNally, Universitat Pompeu Fabra.
2011 – 2012 Spanish government, EXPLORA programme, FFI2010-09464-E ("A distributional semantic model for fully recursive phrasal meaning"), €32,000. PI: Louise McNally, Universitat Pompeu Fabra.
2010 – 2012 Spanish government, TIN2009-14715-C04-04 ("KNOW2: Language understanding technologies for multilingual domain-oriented information access"), €160,600. PI: Jordi Turmo, Universitat Politècnica de Catalunya.
2009 – 2013 Catalan government, Consolidated Research group "GPLN: Grup de Processament del Llenguatge Natural" (2009 SGR 1082). PI: Horacio Rodríguez, Universitat Politècnica de Catalunya.
2008 – 2013 European Union, PASCAL 2: Pattern Analysis, Statistical Modelling, and Computational Learning 2 (EU Network of Excellence). PI: J. Shawe-Taylor, University College London.
2007 – 2010 Spanish government, HUM2007-60599/FILO ("OntoSem: Natural language ontology and the semantic representation of abstract objects"), €109,100. PI: Louise McNally, Universitat Pompeu Fabra.
2006 – 2009 Spanish government, TIN2006-15049-C03-03 ("KNOW: Developing large-scale multilingual technologies for language understanding"), €100,000. PI: Lluís Padró, Universitat Politècnica de Catalunya.
2006 – 2007 Spanish government, international cooperation programme, HA2005-0100 (co-PI: Graham Katz, U. Osnabrück, €11,040) and HF2005-0177 (co-PI: Philip Miller, CNRS-U. Lille, €10,830) ("Natural language ontology for reference to facts and eventualities"). PI: Louise McNally, Universitat Pompeu Fabra.
2005 – 2009 Catalan government, Consolidated Research Group "Unitat de Recerca en Lingüística" (2005SGR00067). PI: Toni Badia, Universitat Pompeu Fabra.
2005 – 2008 Spanish government, grant HUM2004-05321-C02-02 ("ARQUITEXT: Arquitectura integrada para el tratamiento avanzado de textos"), €25,920. PI: Toni Badia, Universitat Pompeu Fabra.
2005 – 2008 Spanish government, HUM2004-04463 ("NOCANDO: Construcciones no canónicas en el discurso oral: estudio transversal y comparativo"), €21,800. PI: Enric Vallduví, Universitat Pompeu Fabra.
2004 – 2007 European Union, IST–FP6-003768 ("METIS-II: Statistical Machine Translation using Monolingual Corpora: from Concept to Implementation"), €1,000,000. General Coordinator: Stella Markantonatou (ILSP, Greece); PI at Universitat Pompeu Fabra: Toni Badia.
2001 – 2004 Spanish government, TIC2000-1681-C02-01 ("PrADo: Sistema de preparación automatizada de documentos"), €36,000. PI: Toni Badia, Universitat Pompeu Fabra.

• SUPERVISION OF GRADUATE AND UNDERGRADUATE STUDENTS

2012 Master's thesis (co-supervision), M. Cornudella Gaya, U. Pompeu Fabra
2010 PhD thesis project (co-supervision), C. Sánchez-Marco, U. Pompeu Fabra
2009 Master's thesis, S. Reese, Institut Supérieur de l'Aéronautique et de l'Espace
2008 – 2009 Undergraduate research assistant, D. Berndt, U. Politècnica de Catalunya

• TEACHING ACTIVITIES

Fall 2014 Lecturer – Morphology, syntax, and introduction to linguistics (B.A., master's), U. Pompeu Fabra, Spain
Spring 2014 Lecturer – Syntax and semantics (B.A.), University of Texas at Austin, US
07/2009 Lecturer – Computational lexical semantics, ESSLLI 2009 *(European Summer School in Logic, Language and Information*), Bordeaux, France
2008 – 2011 Lecturer – Computational linguistics, technology for translation (B. A, master's), U. Pompeu Fabra, Spain
2001 – 2007 Teaching assistant – Computational linguistics, technology for translation (B. A, master's), U. Pompeu Fabra, Spain

• ORGANISATION OF SCIENTIFIC MEETINGS

03/2013 Co-organiser, *IWCS 2013 Workshop: Towards a formal distributional semantics,* Potsdam, Germany
03/2009 Co-organiser, *Jornada del Processament Computacional del Català* (Computational Processing of Catalan Workshop), Barcelona
03/2009 Co-organiser, *Nanoworkshop on statistical physics and linguistic*s, Barcelona
07/2008 Co-organiser, *Coling 2008 workshop on human judgements in Computational Linguistic*s, Manchester, UK

• INSTITUTIONAL RESPONSIBILITIES

2004 – 2006 Doctoral student representative, Department of Translation and Philology Council, U. Pompeu Fabra, Spain
1995 – 1996 Student representative, Department of Spanish Philology Council, U. Autònoma de Barcelona, Spain

• COMMISSIONS OF TRUST

2015 Local co-chair for ESSLLI 2015, Barcelona, Spain, August 3-14 2015.
2014-pres. Editorial Board member, *Linguistic Issues in Language Technologies (LiLT*) journal, beginning September 2014
2013-pres. Information Officer, ACL SIGSEM Board
2013 Area co-chair, **SEM 201*3, Atlanta, US

• REVIEWING / PROGRAMME COMMITTEE MEMBER (excluding workshops)

Journal *Semantics and Pragmatics* (2013), *Natural Language Engineerin*g (2013), *ACM Transactions on Speech and Language Processin*g (2012), *Computational Linguistics* (2012), *Language Resources and Evaluatio*n (2008), *Corpor*a (2008)
Conference ACL 2014, EACL 2014, CoLing 2014, *SEM 2014, IWCS 2013, GL2013, *SEM 2012, LREC 2012, EACL 2012, IJCNLP 2011, ACL-HLT 2011, CoLing 2010, LREC 2010, EMNLP 2009, SEPLN 2009, EACL 2009, ACL 2008
Book *Weak Referentialit*y (eds. A. Aguilar-Guevara, B. Le Bruyn, and J. Zwarts, to be published by John Benjamins)

• CAREER BREAKS IN RESEARCH

29/06/2005 – 31/11/2005 Maternity leave (5 months); Barcelona, Spain
29/11/2007 – 31/05/2007 Maternity leave (6 months); Barcelona, Spain

6. Capacity of the Participating Organisations

Beneficiary: University of Trento (UNITN)	
General Description	The University of Trento (UNITN)[16] serves around 19,000 students and hosts 590 professors and researchers in ten departments and seven research centres. It is only 50 years old, but it has established itself as a centre for excellence in both research and teaching: Among other recognitions, it is considered the best Italian university and number 219 world-wide according to the Times Higher Education Rankings 2013-2014, given its learning environment, research volume and impact, innovation, and international outlook. UNITN has had 118 projects under FP7, out of which 17 under Ideas (ERC Grants), and 46 coordinated by the University. The fellow will be based at CIMeC,[17] the UNITN Center for Mind/Brain Sciences, which studies functional and structural aspects of the mind/brain with an interdisciplinary approach combining methods from Neuroscience, Psychology, Physics, Computer Science, and Artificial Intelligence. CIMeC currently hosts about 40 faculty members, 50 post-docs, and 40 PhD students, organized into five labs, including the Language Interaction and Computation (CLIC) lab, the host of this fellowship. Opened in 2007, CIMeC has quickly been recognized as a center of excellence at the international level, and as such its faculty have been awarded 8 ERC Grants and 12 FP7 projects in total.
Role and Commitment of key persons (supervisor)	Marco Baroni is an associate professor and the director of the CLIC lab (for more information regarding qualifications, see Section 2.3). His current research focuses on two topics which are highly relevant to the present proposal: How to extend distributional semantics to the visual domain, and deriving distributional semantic representations of phrases and sentences with composition techniques. Since the project topics are central to Prof. Baroni's research, he is committed to collaborate with the applicant in the achievement of its goals, and to actively mentor her, in particular to help her attain a deep knowledge of relevant computational and mathematical techniques.
Key Research Facilities, Infrastructure and Equipment	UNITN provides an excellent library and expert administrative support. Especially relevant for the present fellowship are its Scientific Research and Technological Transfer Division unit, which has an office at CIMeC that will directly help grant management, the Knowledge Transfer Office, which provides guidance regarding IPR and knowledge transfer, and the Ethics Committe for ethical aspects. CIMeC offers an international, English-speaking environment with wide experience in hosting both short- and long-term research visits and the highest standards for logistic support: The fellow will have access to an office, all the necessary equipment for the completion of the project (including access to a 20-node cluster server), the centre's IT and administrative support, and help in practical aspects such as housing (see Section 2.3 for details).
Independent research premises?	Yes; UNITN's CIMeC has its own research buildings in the towns of Rovereto (where the fellow with have her own office) and Matarello.
Previous Involvement in Research and Training Programmes	UNITN: for instance, 15 FP6-Marie Curie Actions projects and 7 FP7 – People projects, plus 37 projects funded by the Marie Curie Cofund of Provincia Autonoma di Trento. Of these projects, CIMeC had 3 FP6 Marie Curie fellowships and 2 Marie Curie funded by the Cofund Project of the Provincia Autonoma of Trento.
Current involvement in Research and Training Programmes	UNITN: for instance, 18 FP7-People, 5 projects funded by the Marie Curie Cofund of the Provincia Autonoma of Trento plus 1 Erasmus Mundus Joint Doctorate and 1 Erasmus Mundus Master's in Language and Communication Technologies. Of these, CIMeC is managing 7 ERC grants and 2 FP7-People fellowships and it participates in the Erasmus Mundus Master's programme.
Relevant Publications and/or research/innovation products	F.M. Zanzotto, L. Ferrone and M. Baroni (to appear). When the whole is not greater than the combination of its parts: A decompositional look at compositional distributional semantics. *Computational Linguistics*. M. Baroni, R. Bernardi and R. Zamparelli (2014). Frege in space: A program for compositional distributional semantics. *Linguistic Issues in Language Technologies* 9(6): 5-110. E. Bruni, N. Tram and M. Baroni (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research* 49: 1-47. DISSECT: Toolkit for creating distributional semantic spaces and performing compositional operations: http://clic.cimec.unitn.it/composes/toolkit. SICK: A large, free dataset for the evaluation of compositional semantics: http://clic.cimec.unitn.it/composes/sick.html.

[16] http://www.unitn.it.

[17] http://web.unitn.it/en/cimec.

7. Ethics Issues

Ethics Self-Assessment in Part B

1)

This project will develop a computational system that makes predictions about which entities (in databases and images) are described by a particular linguistic expression. The predictions made by the computational system will be evaluated in various empirical ways, including a set of experiments in which we will use linguistic expressions describing images, collected from native English speakers via an anonymous online survey (see description of Experiment 2 in Section 2.1 and Work Package 2 in Section 4.1). I will gather two types of annotation: A natural description of the image, to gather salient modifiers (*red car*), and a true but non-salient description, like *car with wheels*. The data collected from subjects will be in anonymous format, and a processed dataset in aggregate and anonymous form will be made available to other interested researchers.

I will ensure that the experimental design and data collection and storage complies with the EU legislation on Ethics and national legislation and good practices on research ethics (TC issues are not involved, since the subject recruiting and payment is dealt with by a third party; see part (2) of the self-assessment below for details) by asking for approval of the data collection experiment to the Ethical Committee for the Experimentation with Human Beings (*Comitato Etico per la Sperimentazione con l'Essere Umano)* of the University of Trento (henceforth, UNITN Ethical Committee) and troubleshooting with them as needed.

I will ask for approval for the data collection from the UNITN Ethical Committee upon notice that my proposal has been selected for funding. **I will not carry any data collection until the approval of the UNITN Ethical Committee is in place and a scanned copy of all documents proving compliance with existing EU and national legislation on ethics have been received by the REA.** I expect to have the required documents (opinion from the UNITN Ethical Committee, any other ethics-related documents mandatory under EU or national legislation as advised by the UNITN Ethical Committee) by the date the Grant Agreement is signed. However, note that the experiment with ethical implications will be carried out starting in month 9 of the fellowship (see Section 4.1), such that the development of the project will not be impaired by any delays in the processing of the ethics documents.

2)

OBJECTIVE

The objective of the data collection is to obtain natural descriptions of images, which I will test again computational models to evaluate and analyse them. The data will be used as a benchmark. I will collect the data from healthy adult volunteers, not targeting any vulnerable population.

METHODOLOGY

Data collection procedure. I will collect input from native English speakers via an anonymous online survey. The online survey will be presented to subjects using the *CrowdFlower* crowdsourcing service.[18] This online service puts scientists and software developers in touch with a large pool of subjects that volunteer to participate in online questionnaires and surveys. Using CrowdFlower offers the following advantages:

- A large English-speaking subject pool can be reached;
- Subject recruitment and reimbursement are carried through the CrowdFlower infrastructure;

[18] http://crowdflower.com.

- Subject anonymity is guaranteed by the fact that subjects are recruited and reimbursed by CrowdFlower, without the experimenter's direct involvement;
- CrowdFlower provides a user-friendly interface to design and submit the surveys, and to collect the results.

The volunteers, all legal adults in their countries of residence, visualize a list of active tasks on the CrowdFlower website (or on the websites of associated services), and decide the tasks they want to take part in. In this list, the potential subjects will see the title and a short description of the task. The task will only be visible to subjects that reside in an European English-speaking country (UK and Ireland).

If a subject chooses to take the survey, he/she will be first presented with a page showing an information sheet and request for informed consent (attachment 1). If the subject agrees to take part in the questionnaire, he/she will see a second page showing the information sheet with the personal data processing information and consent form (attachment 2). If the subject proceeds, the next pages will present him/her with the instructions followed by the data collection proper.

Subjects can progress at their own pace and suspend or terminate the task at any time. A single subject will maximally provide 100 image descriptions (to guarantee enough variety in the subject sample, for statistical representativeness purposes). CrowdFlower handles question randomization and other aspects related to the management of the experimental stimuli (survey questions) automatically.

Subjects are reimbursed for the time they devote to the research on the basis of the number of responses they provide, at the rate of 2 Euro-cent per response. I aim to collect at least 10 different descriptions for each image for each of two types of annotation (true-and-salient and true-but-non-salient attributes; see Section 2.1 above), for 500 images, so a total of 10,000 datapoints.

Once data collection is completed, CrowdFlower makes a spreadsheet with the results available to the researchers, on a password-protected page. The spreadsheet contains all the ratings that were collected. Subjects are identified by a numeric ID, and it is possible to reconstruct the geographical location (city) where they report to reside. No other personal information pertaining to the subjects is available to the researchers.

The data will be manually examined and filtered for uncooperative responses or non-native subject participation. Depending on the results of the experiment, either all remaining responses or only responses given by more than one subject will be retained. The resulting noun phrases will be used in the computational modelling experiments as described in Section 2.1.

Please note that, during the project, considerable resources will be devoted to the design and implementation of the surveys. It is thus not possible to present a full description of the final format and full contents at this stage. As far as the contents are concerned, images will involve familiar, everyday objects and scenes. They will be semi-automatically chosen from ImageNet, controlling for the presence of both certain nouns (automatically) and certain modifiers of interest (manually). Potentially contentious or emotionally-charged images will be carefully avoided in the selection of the material. Notice also that we will run one or more pilot studies before launching the data collection, with the same procedure we are describing here, also on CrowdFlower.

Participants. Participants are adult volunteers that have decided to subscribe, as potential subjects, to the CrowdFlower service. Only CrowdFlower subscribers that reside in an European English-speaking country (UK and Ireland) can take part in our surveys. Subjects are moreover requested to take part in the surveys only if they are native speakers of English (although I have no way to enforce this requirement). The subjects will be informed about the possibility of taking part in our study by browsing the list of tasks offered in CrowdFlower and associated services. Since they do not reside in Italy and I will not be able to identify them, it is extremely unlikely that the subjects will hold any direct relationship with members of the research group, or that they will feel any pressure to participate in the study.

Informed consent procedures. Before they can take part in the procedure, subjects must read an information sheet and express their informed consent (see attachment 1). Only subjects who express their consent will be allowed to take the surveys. Subjects are invited to contact me (email address provided on the information sheet) in case they have doubts or require further information. The informed consent form will be verified by the UNITN Ethical Committee to ensure that it complies with Italian and EU legislation.

Privacy and confidentiality. Before they can take part in the procedure, subjects must read a data handling and privacy information sheet and express their consent (see attachment 2). Only subjects who express their consent will be allowed to take the surveys. The data handling and privacy information sheet will be verified by the UNITN Ethical Committee to ensure that it complies with Italian and EU legislation. I will not collect the subject names, nor any personal data (besides the city where subjects declare to reside, an information that is automatically provided to us by CrowdFlower). Each subject will be identified by a unique numeric ID. The data will be stored on the server of the CLIC laboratory of the Center for Mind/Brain Sciences of the University of Trento (server physically located in Rovereto, Italy). The data will be stored in a directory with read/write access permissions granted to my supervisor and me only. System administrators, that will also have access to the directory, are required by Italian law to respect privacy and confidentiality for all data stored on the server. The generated noun phrases and their associated images, without subject IDs, will be shared with other researchers who might request them for other studies with research objectives similar to the present project.

Data storage and retention. The data will be downloaded from the password-protected CrowdFlower result page directly onto the CLIC server. The data will be stored on the CLIC server in Palazzo Fedrigotti (c.so Bettini 31, Rovereto), under the responsibility of the Center for Mind/Brain Sciences system administrator. The server provides redundant data storage and periodic automated backup. Given their scientific value, it is in our interest to preserve the data as long as possible, and in any case for at least 10 years from the end of the research. The participants can ask us at any time to remove their data from the set, as long as they provide us with their ID (because of the anonymous data collection procedure, we do not associate data with names). Only the researchers directly involved in the data collection and analysis procedure will have full access to the data. A database with a set of noun phrases per image, without subjet IDs, will be made freely available for research purposes.

IMPACT

Since I will collect data from healthy adult volunteers in the general population; I will collect no personal data apart from the city of reported residence; and I will not distribute any personal data (not even anonymous subject IDs), I expect no impact issues with the dataset I will create: none of dual use, environmental damage, stigmatisation of particular social groups, political or financial retaliation, or benefit-sharing issues apply. I also do not expect any malevolent use issues to arise, since the data will provide descriptions of everyday objects and scenes.

ATTACHMENT 1: INFORMATION SHEET (draft)

Purpose of the research

- The research aims at creating a database of short object descriptions by native English speakers, in order to come to a better understanding of how human beings use language to communicate, with the possible long term goal of improving human/computer communication.
- You are asked to consent to take part in the data collection survey. The collected data will only be used in anonymous form.

Procedure description

- While filling in the survey [insert description of the task as designed during the project].
- You can work at your own pace and interrupt the task at any time, possibly resuming it later (you can describe a maximum of 100 images).
- The images will depict everyday life objects and scenes, and they will not contain any controversial material (in particular, they do not represent any political opinion, and there is no depiction of violent or sexually explicit scenes).
- We are interested in collecting spontaneous descriptions of images: there is no “right” or “wrong” description.

Potential risk and discomfort

- If, at any time, you should become tired or feel other forms of discomfort, you can simply quit the survey. You are always welcome to come back to it at a later time.

Privacy and data handling

- The experimenters involved in the project will pre-process the data in an anonymous and confidential manner.
- The data we collect, in anonymous format, will be made available to interested researchers on the project web site [insert project web page here].
- Research results will be published in journal articles, conference presentations and via any other mode of scientific exchange and dissemination that will be seen as appropriate by the researchers, while protecting the participants' anonymity.

Voluntary participation and freedom to withdraw from the research

- Your choice to take part in the research is entirely voluntary. You are completely free to choose not to participate, or to withdraw from the study at any moment without any consequence.
- This study was evaluated and approved by the Ethical Committee for the Experimentation with Human Beings (*Comitato Etico per la Sperimentazione con l’Essere Umano*) of the University of Trento.
- The study respects the ethical principles defined by the Helsinki Declaration on research with human beings. If you have doubts on the correct and coherent conduct of the experimentation with respect to what was stated in this form, please contact the University of Trento Dean office, via Belenzani 12, 38100 Trento, Italy.
- For any further doubt or clarification request, do not hesitate to ask! You can contact the principal investigator, Gemma Boleda, by sending an email to the address: [insert UNITN e-mail address that will be provided to me at the start of the fellowship].

INFORMED CONSENT FORM

I HAVE CAREFULLY READ THE INFORMATION SHEET AND I AM FULLY AWARE THAT:

- I am free to withdraw from the survey at any time, without consequences;
- the researchers involved in the project will have access to the data in anonymous format;
- anonymous aggregated data will be downloadable from the project website;
- there will be scientific publications deriving from the research, that will not reveal the identity of the participants;
- the collected data will only be used for research purposes.

HAVING READ, UNDERSTOOD AND ACCEPTED ALL OF THE ABOVE, I *DO/DO NOT* ACCEPT TO TAKE PART IN THE SURVEY. [*DO* and *DO NOT* are hyperlinks that bring the subject either to the next page in the survey, or back to the CrowdFlower home.]

ATTACHMENT #2: INFORMATION AND CONSENT TO PERSONAL DATA PROCESSING (draft)

INFORMATION NOTE IN COMPLIANCE WITH THE SECTION 13 OF THE **LEGISLATIVE DECREE NO. 196/2003** OF THE ITALIAN LAW FOR THE SUBJECTS PARTICIPATING TO THE EXPERIMENTAL PROJECT:

"LOVe: Linking Objects to Vectors in distributional semantics: A framework to anchor corpus-based meaning representations to the external world"

Dear Sir/Madam,

We wish to inform you that according to the LEGISLATIVE DECREE NO. 196/2003 "ITALIAN PERSONAL DATA PROTECTION CODE", everyone has the right to protection of the personal data concerning him or her.

In compliance with the Legislative Decree no. 196/2003 (Italian Personal Data Protection Code), Your personal data will be processed by the experimenters and will be accessed by other authorized researchers according to principles of fairness, lawfulness, transparency and the protection of Your privacy and rights.

We kindly invite you to read carefully the following text, because You will be requested to agree upon an explicit consent to the processing of Your personal data:

In compliance with the section 13 of the Legislative Decree no. 196/2003 (Italian Personal Data Protection Code), we inform You that:

1. the scientific nature of this research requires the acquisition of written survey data, that will be collected only in an anonymous format;
2. the collected data will be kept secure by the experimenters in the forms and with the tools indicated in the following points;
3. data will be processed with electronic devices for the whole period of the experimentation; documentation will be stored in a digital archive at least for the time requested by the existing legislation;
4. data will exclusively be handled for the scientific purposes of the present study;
5. processing, communication and/or dissemination of data collected during the experimentation will be carried out only in anonymous format, under the control and direct responsibility of the research personnel, who will grant the access to the original data only to other researchers officially affiliated to a research institution, under a motivated request in accordance with the purposes of this experimentation;
6. within the limits prescribed by the existing legislation, the members of the research group, other authorized researchers, the members of the Ethical Committee of the University of Trento, and the legislative authorities competent on this subject (e.g. the "Garante" for the Privacy) will be granted direct access to the documentation concerning You, for the purposes of this scientific research, to defend Your rights, and to protect Your privacy;
7. the data collected during this experimentation will be handled by the University of Trento, which is responsible for the data processing, and whose official representative is the Rettore, with address in Trento, via Belenzani n. 12. Dr. Gemma Boleda (email: [insert e-mail]; tel.: [insert tel.]) has been appointed responsible for the data processing. You can refer to the responsible for the data processing to defend Your rights in accordance with the section 7 of the legislative decree n. 196/2003, reported below in its integrity;
8. although Your consent to personal data processing as described above is not mandatory, it is necessary to carry out the experimentation, as well as to fulfil the

Italian law obligations; therefore without Your consent, You will not be able to participate in the experimentation.

7.3.1 Section 7 (Right to Access Personal Data and Other Rights)

1. A data subject shall have the right to obtain confirmation as to whether or not personal data concerning him exist, regardless of their being already recorded, and communication of such data in intelligible form.
2. A data subject shall have the right to be informed
a) of the source of the personal data;
b) of the purposes and methods of the processing;
c) of the logic applied to the processing, if the latter is carried out with the help of electronic means;
d) of the identification data concerning data controller, data processors and the representative designated as per Section 5(2);
e) of the entities or categories of entity to whom or which the personal data may be communicated and who or which may get to know said data in their capacity as designated representative(s) in the State's territory, data processor(s) or person(s) in charge of the processing.
3. A data subject shall have the right to obtain:
a) updating, rectification or, where interested therein, integration of the data;
b) erasure, anonymization or blocking of data that have been processed unlawfully, including data whose retention is unnecessary for the purposes for which they have been collected or subsequently processed;
c) certification to the effect that the operations as per letters a) and b) have been notified, as also related to their contents, to the entities to whom or which the data were communicated or disseminated, unless this requirement proves impossible or involves a manifestly disproportionate effort compared with the right that is to be protected.
4. A data subject shall have the right to object, in whole or in part,
a) on legitimate grounds, to the processing of personal data concerning him/her, even though they are relevant to the purpose of the collection;
b) to the processing of personal data concerning him/her, where it is carried out for the purpose of sending advertising materials or direct selling or else for the performance of market or commercial communication surveys.

CONSENT TO SENSITIVE DATA PROCESSING

After reading and understanding the information note above concerning the research: "LOVe: Linking Objects to Vectors in distributional semantics: A framework to anchor corpus-based meaning representations to the external world"

- I agree
- I do not agree

to the processing of my sensitive data within the present research, in accordance with the terms and conditions mentioned the points1-8 above. I also explicitly authorize the members of the research group, other authorized researchers, the members of the Ethical Committee of the University of Trento, and the legislative authorities competent on this subject (that must respect data confidentiality) to access and consult my data. The processing of the data collected in this experimentation, as well as their communication to other subjects and/or their dissemination for scientific purposes, are authorized under the direct control and responsibility of the experimenter. [The *I agree* and *I do not agree* bullet points are hyperlinks that bring the subject either to the next page in the survey, or back to the CrowdFlower home.]

ENDPAGE

MARIE SKŁODOWSKA-CURIE ACTIONS

Individual Fellowships (IF)
Call: H2020-MSCA-IF-2014

PART B

"LOVe"

"**L**inking **O**bjects to **Ve**ctors in distributional semantics: A framework to anchor corpus-based meaning representations to the external world"

This proposal is to be evaluated as:

[Standard EF]