European Research Council

ERC Starting Grant
Research proposal (Part B section 1 (B1))

Compositional Operations in Semantic Space

COMPOSES

Cover Page:
- Principal Investigator (PI): Marco Baroni
- Host institution: Center for Mind Brain/Sciences (CIMeC) of the University of Trento (UNITN)
- Proposal full title: Compositional Operations in Semantic Space
- Proposal short name: COMPOSES
- Proposal duration in months: 60

"Pink dogs are rare". You understood this sentence even if you've never read it before, because you know the meanings of thousands of words (including "*pink"*, "*dogs"* and "*rare"*) and how to construct the meaning of a novel sentence from the meanings of its parts. For decades, scientists in different fields have tried to develop computational systems that "understand" sentences as humans do. They have, however, failed either the challenge of *coverage* (knowing thousands of words) or that of *compositionality* (putting together the parts to reconstruct the meaning of new sentences). The goal of COMPOSES is to make a breakthrough on both fronts using powerful novel tools made possible by the existence of the Web. On the coverage front, I have developed the largest linguistically processed English corpus in existence, from which I have extracted a very large set of word-level semantic representations, based on both statistical and linguistic properties of word distribution. On the compositionality front, building on my theoretical linguistics background and collaborating with experts in formal semantics, I have prototyped a method to automatically build phrase representations out of corpus-extracted word representations and corpus-learned composition functions. In COMPOSES, I will scale the approach up to the sentence level, testing it extensively on tasks that tap into semantic knowledge, in order to show that its performance is significantly human-like in many important respects. Thanks to the novel system that I propose and the human-derived semantic data sets that I will collect and make publicly available, COMPOSES will have a large impact on theoretical and computational linguistics, cognitive science and artificial intelligence, paving the way for a new generation of computational systems endowed with a human-like capacity to produce and understand the meaning of full utterances.

Section 1: *The Principal Investigator*

1(a) Scientific Leadership Potential

I obtained both my MA and PhD degrees from the UCLA Department of Linguistics, where I acquired a strong theoretical linguistics background. Since the very beginning, I have been interested in broadening the methodological boundaries of this field by taking inspiration from disciplines that study language from other perspectives: in particular, computational simulations from computer science and experimental methods to elicit and analyze language data from cognitive science.

After the completion of my education, I worked at an American company specializing in speech synthesis and recognition and at the Austrian Research Institute for Artificial Intelligence. These experiences made me aware, on the one hand, of the enormous potential of corpus-driven statistical learning models applied to language, and, on the other, of the need to construct large-scale linguistic resources to extract better statistical models. Thus, soon after taking a tenured researcher position at the University of Bologna, I became one of the main instigators and coordinators of the nascent community of linguists interested in creating large linguistic resources from the Web (the WaCky community) and I founded the Web as Corpus special interest group of the Association for Computational Linguistics, of which I am secretary. My most significant contribution in this area has been the construction of the largest publicly available linguistically annotated corpora for English, German, French and Italian (downloaded more than 600 times as of October 2010, by a wide variety of institutions). My corpus construction work has received extensive recognition, including a best paper award at the LREC 2008 Web-as-Corpus workshop, an invited plenary talk at an international conference, five invited talks at various institutions and workshops, a University of Bologna Marco Polo grant, a national Italian FIRB grant and requests to review three NSF grant proposals.

Following my relocation to the fertile environment of the Center for Mind/Brain Sciences of the University of Trento, I set out to use the English corpus I created to tackle one of the most ambitious goals of linguistics: devising a computational system that extracts linguistic meaning from data. I developed an innovative computational semantic model that induces concept representations in terms of cognitively plausible properties and applies the knowledge thus harvested to semantic tasks in a flexible human-like manner. This work brought computational corpus-based semantics considerably closer to cognitive science and theoretical linguistics, resulting in two recent articles in high-impact journals with interdisciplinary focus (Cognitive Science 2010, Computational Linguistics to appear). General recognition for my work on this front includes a Google Research Award, a national Italian PRIN grant, an invited plenary talk at an international conference, invited talks at three workshop, two invited lecture series, an Invited Scholar Fellowship from the National Institute for Japanese Language, an invitation to act as external consultant in a nationally funded Spanish project and a contract with Cambridge University Press to write an introductory volume on corpus-based semantics.

As a linguist, my work with computational simulations is motivated by my interest in what they tell us about human linguistic competence. Consequently, I have also a strong interest in collecting and analyzing subject-elicited language data, in order to construct a cognitively sound empirical basis to compare computational models against. I have proposed innovative ways to collect linguistic data (ranging from articulatory phonetics to Web surveys, to neuro-imaging, to crowdsourcing and "games-with-a-purpose") and data analysis techniques involving advanced statistical ideas, such as mixed linear models. Recognition in this area includes two invited methodological chapters in the Mouton de Gruyter HSK series *Handbook of Corpus Linguistics,* two invited methodological chapters to appear in the *Enciclopedia Treccani* of Italian and three invited statistics/methodological lecture series.

Thanks to the resources I developed and the uniquely varied background I built until now, I am confident that I can be the first person to take the next step the field needs: merging corpus-driven computational modeling, insights from theoretical linguistics and experimental methods in order to develop a semantic system that achieves large coverage, meets the compositionality challenge and is supported by strong cognitive evidence.

Until now, my work has been fragmented into a series of short term projects supported by small funds or voluntary work. COMPOSES will finally give me the breadth, as a true "consolidator", to pool the knowledge, experience and ideas that I have assembled into a large-scale project that will impact all the relevant fields.

1(b) Curriculum Vitae
Constantly updated information at: http://clic.cimec.unitn.it/marco

Education

- **PhD** in Linguistics, University of California, Los Angeles, June 2000. Dissertation Title: Distributional cues in morpheme discovery: A computational model and empirical evidence. Supervisor: B. Hayes. Main results were reported in single-author peer-reviewed article in *Yearbook of Morphology 2003*.
- **MA** in Linguistics, University of California, Los Angeles. December 1997. Thesis Title: The representation of prefixed forms in the Italian lexicon: Evidence from the distribution of intervocalic [s] and [z]. Supervisor: B. Hayes. Main results were reported in single-author peer-reviewed article in *Yearbook of Morphology 1999*.
- Italian "**Laurea**" degree (approximately equivalent to MA), *summa cum laude*, University of Padua, April 1995. Thesis Title: La relazione tra struttura segmentale e costituenza moraica [The relation between segmental structure and moraic constituency]. Supervisors: A. Mioni and L. Vanelli. Work related to thesis resulted in single-author peer-reviewed articles in *Rivista di Grammatica Generativa* (1993) and *Romance Linguistics and Literature Review* (1994).

Post-doctoral employment history

- November 2006 - present: **Researcher** (tenured position), CLIC laboratory, Center for Mind/Brain Sciences (CIMeC), **University of Trento**, Italy.
- October 2002 - October 2006: **Researcher** (tenured from 2005), Advanced School for Interpreters and Translators, **University of Bologna**, Italy.
- September 2001 - August 2002: **Post-doctoral researcher** (position funded by EU R&D project FASTY), **Austrian Research Institute for Artificial Intelligence** (ÖFAI), Vienna, Austria.
- July 2000 - August 2001: **Computational Linguist**, **Conversay Conversational Corporation**, Redmond WA, USA.

Teaching and supervision at the University of Trento
Updated information on my teaching activities at: http://people.lett.unitn.it/baroni/

- **Main proponent** and **coordinator** of BA and MA curricula in Humanities and Quantitative/Formal Methods at the School of Letters and Philosophy.
- **Teaching duties**: BA-level course on introductory programming for text analysis, MA/MSc-level course on text processing, doctoral level course on statistics.
- **PhD supervision**: 1 completed (2010); 2 students expected to defend in early 2011; 1 student and 1 co-supervised student starting in 2010/2011. I co-supervise 1 student to work on content word composition in corpus-based semantics: we expect the student to collaborate closely with the COMPOSES team.
- **Bachelor and Master supervision**: 2 BA theses completed (2009, 2010); 1 MA thesis completed (2010); 3 students working on MA/MSc theses (all expected 2011).

Grants (Funding ID 1)

- **Semantic spaces from text and images**, Google Research Award, 2010-2013 (funds one doctoral student): project aims at extending current text-based semantic models with information extracted from automated image analysis. Like COMPOSES, this project will advance theoretical and empirical aspects of computational semantics. There will be an overlap of about 2 years, when the doctoral student funded by this project will collaborate closely with the COMPOSES team on the development of tools, data and test sets.
- **Conceptual representation in the blind: Empirical data and computational simulations**, Italian PRIN project, 2010-2012 (PI of Trento unit; will fund one 1-year post-MSc position in 2011): elicitation of concept descriptions from congenitally blind and sighted subjects, and comparison with concept properties generated by corpus-based semantic models. Experimental and analytical methodologies developed within the project will be exploited in COMPOSES. Trento contribution should be nearly completed by the time COMPOSES starts.
- **PAISA'**, Italian FIRB project, 2009-2011, (PI of Trento unit; funded one 1-year post-doc and one 1-year post-MA positions): building a large and representative Web-derived corpus of Italian from

free-license text. Trento contribution completed, no temporal overlap with COMPOSES.

- **LiveMemories**, Trentino Province PAT project, 2008-2011 (I am not PI, but I contributed to proposal writeup, and was assigned supervision of a 3-year post-MSc researcher): new technologies for sharing, organizing and searching collective memories. No temporal overlap with COMPOSES.

Selected publicly available tools and resources

- **WaCky** (with Silvia Bernardini and others): huge linguistically annotated corpora for multiple languages
- **DM** (with Alessandro Lenci): precompiled corpus-based semantic model and utilities
- **Semantic norms** for German and Italian (with Gerhard Kremer)
- **zipfR** (with Stefan Evert): a toolkit for lexical statistics in R
- **BootCaT** (with Silvia Bernardini): a toolkit for bootstrapping corpora and terms from the Web
- **Morph-it!** (with Eros Zanchetta): a free Italian morphological lexicon
- **La Repubblica corpus** (with Silvia Bernardini and others): a large corpus of Italian newspaper text

Other activities

- **Workshop (co-)organization**: GEMS 2010 (submitted), ESSLI 2008 Distributional Lexical Semantics (Hamburg), Contextual Information in Semantic Space Models at Context 2007 (Roskilde), Web as Corpus 1 (2005, Forlì), 2 (2005, Birmingham) and 3 (2006, Trento)
- The Italian part-of-speech tagger developed by my team was ranked second best in the **EVALITA 2007 evaluation campaign**
- Co-organized the first **CLEANEVAL shared task** for Web page cleaning (2007)
- Co-founder and secretary of the **Special Interest Group** of the Association for Computational Linguistics (ACL) **on Web as Corpus**
- **ESSLLI 2006 course** instructor (with Stefan Evert): Counting words: an introduction to lexical statistics (Malaga)
- I maintain, with Stefan Evert, **SIGIL**, an **online introduction to statistics for linguists**
- In **program committee** of more than 10 international conferences (including ACL, EACL, COLING, IWCS, EMNLP – best reviewer award at EMNLP 2010) and more than 15 international workshops
- **Reviewer** for more than 15 **journals** (including Natural Language Engineering, IEEE Intelligent Systems, Language Resources and Evaluation Journal, Cognitive Linguistics, Europhysics Letters, Artificial Intelligence Journal, Morphology and the Journal of the Acoustical Society of America) and 2 **books**
- **Reviewer** for several **funding agencies**, including the US National Science Foundation and the UK Economic and Social Research Council

COMPOSES commitment (Funding ID 2)

During the COMPOSES period, I am fully committed to the project, except for a maximum of 80 hours of teaching per year and doctoral student supervision (the 2 doctoral students outside the project will work on related themes). I will not apply to funding for other projects that would overlap with the COMPOSES period. COMPOSES will also constitute the core activity of the CLIC laboratory for the period of the project.

1(c) Early Achievement-Track-Record

Recognition of achievements

Awards

- Google Research Award (2010)
- Best paper award (as co-author), Web as Corpus 4 Workshop at LREC (2008)
- Invited Scholar Fellowship, the National Institute for Japanese Language (2007)
- Marco Polo Scholarship for a research period abroad, University of Bologna (2005)
- Chancellor Fellowship, University of California, Los Angeles (1995-2000)

Invited presentations to international conferences and international advanced schools

M. **Baroni**. 2010. Web 2.0 as corpus: One decade of textual analysis with Web data. Invited plenary talk at *JADT 2010*, Rome (Italy).

M. **Baroni**. 2009. Vector-based models of semantic relatedness. Invited mobility program course at *GLIF*, Pompeu Fabra University, Barcelona (Spain).

M. **Baroni**. 2009. Statistical methods for lexical acquisition. Invited lecture series at the *TRIPLE Winter School on the Lexicon*, Rome (Italy).

M. **Baroni**. 2008. Distributional semantics: From ad hoc solutions to persistent models. Invited plenary talk at *IS-LTC 2008*, Ljubljana (Slovenia).

M. **Baroni**. 2008. Distributional semantics. Invited lecture at the *Beyond Short Units* workshop of the "Sound to Sense" Marie Curie Research Training Network, Naples (Italy).

M. **Baroni** and S. Evert. 2007. Statistical programming in R for computational linguists. Invited course at the *Computational Linguistics Fall School* of the German Linguistics Association, Potsdam (Germany).

Selected publications

Citation counts from Google Scholar (Oct 2010), self-citations manually removed. Complete publication list **(**full text of most articles available) at: http://clic.cimec.unitn.it/marco/research.html

Selected publications in leading peer-reviewed journals and selected book chapters

Journal rank quartile from ISI Web-of-Science Journal Citation Reports if available; in all other cases, journal category from European Science Foundation ERIH Initial List: Linguistics (2007)

M. **Baroni** and A. Lenci. To appear. Distributional Memory: A general framework for corpus-based semantics. *Computational Linguistics*. Journal ranked in Q1 of *Linguistics* and *Computer Science, Interdisciplinary Applications*. Citations: 1. Here and in the Cognitive Science paper below, I introduce a new corpus-based semantic model of word meaning that adapts flexibly to multiple semantic tasks and shares interesting properties with human semantic cognition. The model and its extensive evaluation work reported in these articles will constitute starting points for COMPOSES model and evaluation.

G. Kremer and M. **Baroni**. To appear. A set of semantic norms for German and Italian. *Behavior Research Methods*. Journal ranked in Q1 of *Psychology, experimental* and *Psychology, mathematical.* Citations: 0.

M. **Baroni**, B. Murphy, E. Barbu and M. Poesio. 2010. Strudel: A corpus-based semantic model based on properties and types. *Cognitive Science* 34 (2): 222-254. Journal ranked in Q1 of *Psychology, experimental*. Citations: 9.

V. Pirrelli, E. Guevara and M. **Baroni**. 2010. Computational issues in compound parsing. In S. Scalise and I. Vogel (eds.), *Cross-disciplinary issues in compounding*, Amsterdam: Benjamins: 271-286. Citations: 0.

M. **Baroni**, S. Bernardini, A. Ferraresi and E. Zanchetta. 2009. The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora. J*ournal of Language Resources and Evaluation* 43 (3): 209-226. ERIH B category (formerly *Computers and the Humanities*, ERIH A category). Citations: 28. The corpus construction and annotation work described here has recently been extended with a full dependency parse of the English corpus and of the English Wikipedia. The resulting enlarged corpus will constitute the main data source of COMPOSES.

M. **Baroni**, E. Guevara and R. Zamparelli. 2009. The dual nature of deverbal nominal constructions: Evidence from acceptability ratings and corpus analysis. *Corpus Linguistics and Linguistic Theory* 5 (1): 27-60. ERIH C category. Citations: 1. We study a linguistic problem by combining corpus data, Web-collected graded linguistic judgments and advanced statistical analysis (mixed effect linear models). We will apply similar elicitation and analysis techniques in COMPOSES.

M. **Baroni**. 2009. Distributions in text. In A. Lüdeling and M. Kytö (eds.), *Corpus linguistics: An international handbook* (HSK series), Volume 2, Berlin: Mouton de Gruyter: 803-821. Citations: 22.

M. **Baroni** and S. Evert. 2009. Statistical methods for corpus exploitation. In A. Lüdeling and M. Kytö (eds.), Corpus linguistics: *An international handbook* (HSK series), Volume 2, Berlin: Mouton de Gruyter: 777-802. Citations: 5.

M. **Baroni** and A. Lenci. 2008. Concepts and properties in word spaces. *Italian Journal of Linguistics* 20(1)): 55-88. ERIH B category. Citations: 11.

M. **Baroni** and S. Bernardini. 2006. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing* 21(3). 259-274. ERIH B category. Citations: 24.

H. Trost, J. Matiasek and M. **Baroni**. 2005. The language component of the FASTY text prediction system. *Applied Artificial Intelligence* 19(8). 743-781. Journal ranked in Q3 of *Engineering, Electrical & Electronic*. Citations: 17.

M. **Baroni**. 2003. Distribution-driven morpheme discovery: A computational/experimental study. In G. Booij and J. van Marle (eds.), *Yearbook of Morphology 2003*, Dordrecht: Springer. 213-248. Citations: 20.

M. **Baroni**. 2001. The representation of prefixed forms in the Italian lexicon: Evidence from the distribution of intervocalic [s] and [z] in Northern Italian. In G. Booij and J. van Marle (eds.), Y*earbook of Morphology 1999*, Dordrecht: Springer. 121-152. Citations: 10.

M. **Baroni** and L. Vanelli. 2000. The relationship between vowel length and consonantal voicing in Friulian. In L. Repetti (ed.), *Phonological theory and the dialects of Italy*. Amsterdam: John Benjamins. 13-44. Citations: 14.

Most cited publications in peer-reviewed conference proceedings

Papers cited more than 30 times as of Oct 2010, ordered by citation count.

M. **Baroni** and S. Bernardini. 2004. BootCaT: Bootstrapping corpora and terms from the web. *Proceedings of LREC 2004*, Lisbon: ELDA. 1313-1316. Citations: 116.

M. **Baroni**, J. Matiasek and H. Trost. 2002. Wordform- and class-based prediction of the components of German nominal compounds in an AAC system. *Proceedings of COLING 2002*, East Stroudsburg PA: ACL. 57-63. Citations: 59.

M. **Baroni** and A. Kilgarriff. 2006. Large linguistically-processed Web corpora for multiple languages. *Conference Companion of EACL 2006*. 87-90. Citations: 56.

P. Keating, M. **Baroni**, S. Mattys, R. Scarborough, A. Alwan, E. Auer and L. Bernstein. 2003. Optical phonetics and visual perception of lexical and phrasal stress in English. *Proceedings of the 15th International Congress of Phonetic Sciences*. 2071-2074. Citations: 44.

A. Ferraresi, E. Zanchetta, M. **Baroni** and S. Bernardini. 2008. Introducing and evaluating ukWaC, a very large Web-derived corpus of English. *Proceedings of the WAC4 Workshop at LREC 2008*. **Best paper award**. Citations: 37

M. **Baroni** and S. Bisi. 2004. Using cooccurrence statistics and the web to discover synonyms in a technical language. *Proceedings of LREC 2004*. Citations: 32.

Books (as author and edited)

M. **Baroni** and A. Lenci. In preparation. *Distributional semantics*. Cambridge University Press (Contract signed in April 2010, manuscript due in 2011; will be first volume in new CUP series on linguistics and natural language processing).

M. **Baroni**, S. Evert and A. Lenci (editors). 2008. *Bridging the gap between semantic theory and computational simulations*. ESSLLI workshop proceedings. Hamburg: FOLLI.

M. **Baroni**, A. Lenci, and M. Sahlgren (editors). 2007. *Beyond words and documents*. CONTEXT workshop proceedings. Roskilde: Roskilde University

M. **Baroni** and S. Bernardini (editors). 2006. *Wacky! Working papers on the Web as Corpus*, Bologna: Gedit.

A. Kilgarriff and M. **Baroni** (editors). 2006. *Proceedings of the 2nd International Workshop on the Web as Corpus*, East Stroudsburg PA: ACL.

Other publications of high relevance to COMPOSES

M. **Baroni** and R. Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. *Proceedings of EMNLP 2010*. 1183-1193. Pilot study applying the COMPOSES approach to adjective-noun combinations, with promising preliminary results.

A. Herdağdelen and M. Baroni. Submitted. Stereotypical gender actions can be extracted from Web text. Our first foray into crowdsourcing with CrowdFlower. We will use this data collection extensively within COMPOSES.

Section 1d: *Extended Synopsis of the project proposal*

Background and goals

Semantics is the cognitive faculty that allows us to rely on language to reason and communicate about states of the world and of our minds. As such, it constitutes the interface of language with conceptual knowledge and other cognitive faculties. Researchers in linguistics, cognitive science and artificial intelligence have tried, for decades, to devise an artificial system endowed with human-like capabilities to use natural language semantics. From a theoretical perspective, computational models of semantics (especially models learning from data) are of enormous interest because they can shed light on how humans themselves acquire and use semantics. Computational models of semantics are also at the center of applied research, since any "intelligent" interaction with machines requires sharing semantic knowledge (question answering, information retrieval and machine translation are among the applied domains that have recently benefitted from progress in data-induced semantic models: Jurafsky and Martin 2008, Manning et al, 2008). The desiderata for a fully satisfactory computational semantic system include:

- Being equipped with semantic representations for thousands of words (at end of high-school, an average Western person might know the meaning of as many as 60,000 words, Aitchison 1993);
- Possessing composition rules that combine these words to construct/interpret sentences that have not been encountered before (you know what "*pink dogs are rare*" means even if this is the first time you read it);
- Agreeing with human behaviour and intuition on a variety of tasks that tap into semantic knowledge (if humans find *"lions have a mane"* more natural than *"lions are males"*, objective statistics notwithstanding, our system should also do so).

COMPOSES will be the first system that satisfies the three desiderata, at least for an explicitly delimited set of possible English constructions.

Considerable advances have recently been made on methods to acquire semantic representations of single words from large text collections (corpora). *Distributional semantics models* (DSMs), abiding by the Firthian maxim that "You shall know a word by the company it keeps", construct large-scale semantic representations of word meaning by collecting vectors that keep track of patterns of co-occurrence of words in corpora (Turney and Pantel 2010). Angular distance and other linear algebraic properties in the resulting distributional vector space have been shown to correlate with human semantic intuitions. Recent DSMs, and in particular the models I proposed (Baroni et al. 2010; Baroni and Lenci to appear) are becoming increasingly more linguistically sophisticated, and can tackle an impressive array of lexical semantic tasks, from detecting synonyms to concept categorization, to finding prototypical properties of concepts to predicting semantic priming. DSMs, however, do not scale up to full utterances, because they do not yet account for *compositionality* (Frege 1892), that essential and uniquely human property by which we can produce an infinity of meanings by combining words into larger constituents. Without compositionality, corpus-based distributional semantics is of little use, since human communication is based on utterances, not on isolated words. There have been some recent attempts to overcome this limitation (Mitchell and Lapata 2010), but they are restricted to combinations of two *content words* (words with full "lexical meaning" such as adjectives and nouns).

Formal semantics, the research program in theoretical linguistics harking back to the seminal work by Montague (1970), has given a lot of emphasis to compositionality, focusing in particular on how meaning construction is guided by syntactic structure and how certain words/phrases (*grammatical words* such as determiners or modals, but also adjectives and verbs) are, at least in some contexts, *incomplete expressions*, acting as functions when composed with other expressions (Chierchia and McConnell-Ginet 2000). Formal semantics has furthered our understanding of the functional scaffolding that keeps sentences together, but it has largely ignored (with important exceptions, such as Pustejovsky 1995) the problem of capturing the meaning of single words. Moreover, formal semantics has been largely confined to a few "paper and pencil" case studies, since it has not come up with a method to induce semantic representations on a large scale.

A third approach relies on *supervised machine learning* techniques to induce various aspects of semantics, such as semantic roles of verb arguments (Jurafsky and Martin 2008, chapters 19 and 20). In this paradigm, a statistical algorithm extracts generalizations from a corpus of appropriately encoded examples that are labeled with the intended output, and then applies the learned generalizations to new examples. Machine learning provides powerful methods to learn arbitrary functions from example data, where the output of the function might be a discrete *classification* choice or, in the *regression* setting, continuos values

(Hastie et al. 2009). However, each semantic/linguistic phenomenon is modeled separately, and each requires its own hand-coded training examples.

Various empirical benchmarks to evaluate semantic models against have been proposed. Most relevantly for us, there is an extensive literature encompassing cognitive science, information retrieval and computational linguistics on collecting and analyzing *semantic similarity judgments* (see, among many others: Medin et al. 1993, McDonald 2000, and references there). There is also a small but very interesting literature on systematically eliciting graded *plausibility judgments* about sentences (e.g., Bard et al., 1996, Cowart 1997, Weskott and Fanselow 2008). Such judgments have also been harvested using Web surveying techniques (Baroni, Guevara and Zamparelli, 2009), and very recent work suggests that they can be reliably collected on a very large scale via Web-based crowdsourcing services (Munro et al. 2010). Another approach to evaluating semantic models, closer to the formal semantics tradition but recently popular in computational linguistics – where it is known as *recognizing textual entailment* (Dagan et al. 2009) – requires the models to identify various kinds of logical relations holding between sentence pairs (sentence *A* implies/contradicts sentence *B*, etc.).

The COMPOSES project brings together ideas from all the research lines sketched above to develop a novel system that can acquire and process composite meanings. The proposed system is *large-scale* and *data-driven* (unlike formal semantic models), automatically acquiring representations for thousands of words from corpora; it produces representations for *full sentences* using compositional operations (unlike DSMs); and, unlike most machine learning approaches, it is *general-purpose* and it performs function learning *without manually labeled examples*. Two decades of distributional semantics studies have shown that word meaning is approximated, at some level, by a distributional vector. I intend to prove that *sentential meaning is also, at some level, a distributional vector, but one that can be obtained compositionally*. This novel idea will require extensive empirical validation. Thus, the second major component of COMPOSES deals with designing new semantic tasks (involving similarity, plausibility and entailment) to assess the usefulness of the vector-based representation of sentences, and in particular whether COMPOSES-generated sentence representations predict human semantic intuitions. Summarizing, the COMPOSES proposal has the following **ground-breaking aspects**:

1. It is the first large-scale, data-driven model of sentence meaning in English, and more specifically it proposes
2. a vectorial compositional semantics model for sentences, coupled with
3. new tasks to evaluate computational models of sentential semantics.

The COMPOSES approach to compositional distributional semantics

In distributional semantics, the meaning of *all* words is given by vectors derived from contextual information directly extracted from the corpus. I believe that this is the true pitfall of current approaches to compositionality in DSMs: They first build distributional vectors for all words, then combine them. Distributional vectors for grammatical words (or even frequent content words) will tend to have flat distributions over texts (since they occur in all sorts of contexts) and they will thus not be very informative (think of a distributional vector for the "*a"* determiner). Moreover, grammatical words and content words that operate like functions in meaning composition (such as attributive adjectives) will receive a single representation across contexts, independently of the words they compose with. This is also highly problematic – think of all the context-dependent senses that even a relatively contentful preposition like "*in"* can have.

In COMPOSES, I propose instead that, mirroring the distinction between complete and incomplete expressions in formal semantics, there are two main classes of words: those whose meaning is represented by corpus-derived distributional vectors, and those whose meaning is given by *distributional functions*. The representation of these latter words is constructed from the distribution of the argument(s) they take and the constituent they build. An explicit syntax-driven grammar guides the composition process, determining which functions are applied to which arguments, and in which order. The functions are induced from corpus data by collecting distributional vectors representing their inputs and outputs, and using input/output vector pairs to train a regression algorithm. This algorithm produces continuous values that approximate those in the output vectors.

Earlier "compositional" DSMs (if they were applied to composition with a grammatical word such as "*some"*) would need to collect a vector for "*some"* and a vector for, say, "*cat"*, then combining the two to obtain a vector for "*some cat"*. I propose to collect instead vectors from the corpus occurrences of nouns like "*people", "butter", "dog"* and the corresponding phrases "*some people", "some butter", "some dog"*. We

can then use vector pairs such as <"*people*", "*some people*"> to learn a function that transforms the input vectors into vectors that look as much as possible as the corpus-observed output vectors. Importantly, the examples used to train the function can be automatically extracted from corpus data, and they require no manual labeling. The function thus learned is then applied to arbitrary nouns to generate composite vectors. The vector representing "*some cat*" is thus the result of applying the corpus-trained "*some*" function to the corpus-extracted "*cat*" vector. There is no need to collect distributional vectors for grammatical words or other functional elements (we collect vectors for phrases like "*some dog*", not for "*some*"!). The induced functions, having been trained on specific input-output instances and operating on many dimensions, can adjust to the specific input they apply to, implicitly capturing vagueness, context-dependence and polysemy. The "*some*" function applied to "*cat*", for example, will produce a vector that is more like the one of "*some dog*" than the one of "*some butter*". The system is syntax-driven (syntax determines the order of combination) and syntax-sensitive (the same functional element triggers different functions when occurring in different constructions). Importantly, functions can be applied to the output of other functions (or to their own output), building larger constituents up to full sentences. The COMPOSES approach is schematically illustrated in Fig. 1 (the specific composition rules implied by the figure are for illustrative purposes only, and they might be changed in the actual COMPOSES grammar: see the technical discussion in Part B2, and especially the distinction between G1.0 and G2.0).

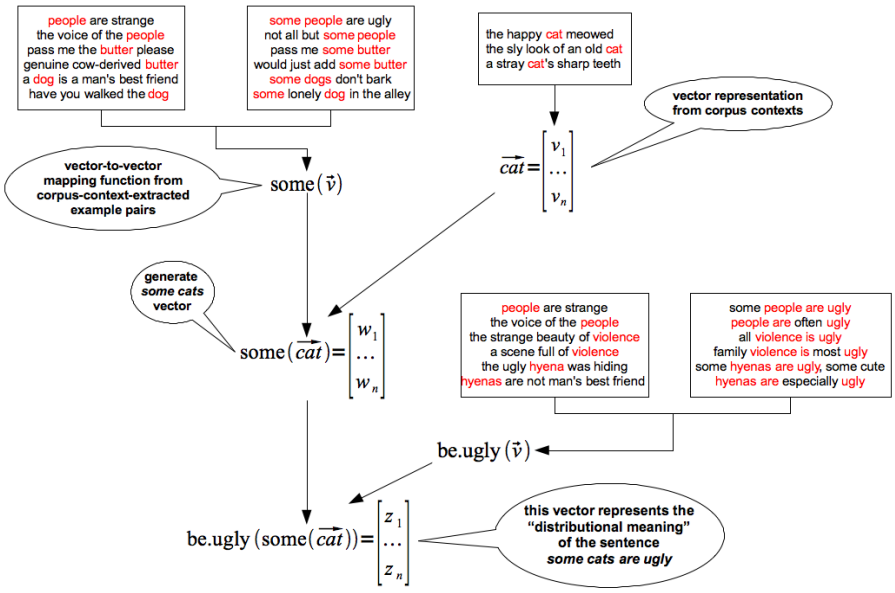

Figure 1: COMPOSES construction of the distributional meaning of the sentence "Some cats are ugly"

Methodology

Work Package 1: Data, tools and infrastructure

Activity 1 (Computational infrastructure) will set up a powerful hardware and software architecture (using the MapReduce distributed computing paradigm). Activity 2 (Data and tool packaging and release) insures that the data and tools released by the project are appropriately engineered and documented to maximize their usefulness to other researchers.

Work Package 2: Computational modeling

Activity 1 (Composition grammar) will specify, maintain and extend the grammar governing the subset of composition rules we intend to model (a first sketch of the grammar is presented in Part B2). The subset covers a sufficiently varied typology of linguistic phenomena to be of practical utility in handling real life utterances. Moreover, while the grammar is restricted in order to keep its implementation and testing feasible, it is instantiated in a distributional semantic space containing representations of thousands of words, so that it has a very large coverage of natural English sentences (the structure of these sentences is constrained by the composition grammar, but their content is virtually unlimited because of the huge number of words and phrases that can fill each grammar slot). Activity 2 (Distributional space construction) builds on my previous work on very large linguistically annotated corpora (Baroni et al. 2009, LREJ) and linguistically sophisticated DSMs (Baroni and Lenci to appear, Baroni et al. 2010) to explore distributional spaces for the representation of simple and composite expressions that capture contexts in all their linguistic

richness (not only lexical collocates, but syntactic and morphological information and shallow discourse structure cues). Activity 3 (Learning distributional functions) implements the abstract linguistic operations spelled out in the Activity 1 composition grammar into concrete functions trained on and applied to objects that live in the Activity 2 distributional space. Following our preliminary work (Baroni and Zamparelli 2010), we will start by assuming that distributional functions are linear and estimate them using *partial least-squares regression* (Hastie et al. 2009, Sec. 3.4). COMPOSES will explore new generalization methods to learn functions that involve constituents high in the syntactic hierarchy (e.g., the predicate functions mapping subject noun phrases onto sentences), where distributional data will be scant even in a distributional space harvested from a huge corpus. We will look in particular at graphical models (Koller and Friedman 2009) as a general framework to develop multi-level learning methods that capture generalizations across both training examples and functions.

Work Package 3: Semantic tasks for empirical evaluation
Given the innovative representation of sentences as vectors, new semantic tasks that illustrate the general usefulness and cognitive soundness of this representation must be developed. COMPOSES can rely here on my previous experience in systematic elicitation of linguistic judgments and statistical techniques to analyze them (see in particular Baroni et al. 2009, CLLT). Experimental data will be collected on a large scale via crowdsourcing services. In Activity 1 (Plausibility ratings), we will collect plausibility ratings by adapting elicitation techniques from experimental linguistics, and tapping into various potentially interesting factors that affect plausibility, such as generic validity of statements ("*some dogs are crippled"* vs. "*all dogs are crippled"*), contradiction ("*mortal beings are immortal"*) and conceptual anomaly ("*talkative tables are eatable"*). Activity 2 (Semantic similarity ratings) extends the existing work on semantic similarity of words and concepts by collecting similarity ratings for full sentences, investigating questions such as: How does inserting/removing modifiers affect similarity between a pair of sentences? How does the presence of entailment or contradiction affect the judgment? How do polysemous words affect sentence similarity in different contexts? In Activity 3 (Entailment), semantics experts will build a data set that illustrates various entailment relations between sentences, manipulating inferential properties such as direction of implication of quantified NPs ("*all dogs"* implies "*some dogs"*) and lexical inclusion ("*dogs"* implies "*animal"*).

In Activity 4 (Model validation), we will use the semantic data sets constructed in the other activities, as well as standard lexical semantic sets (Baroni and Lenci to appear), in order to extensively test the performance of the COMPOSES models and alternatives from the literature, where possible.

Work Package 4: Coordination and dissemination
Activity 1 (Project management) monitors progress and communication flows. It also organizes periodic project reviews by the 3 external consultants, that are world-wide leading scholars in formal and/or computational semantics (Nicholas Asher, Stephen Clark and Katrin Erk). Activity 2 (Dissemination) insures that enough project time is dedicated to the dissemination of project results through an informative website where we will release all data sets, tools and publications produced by the project, as well as by maintaining a systematic presentation and publication schedule and organizing an end-of-project workshop).

Feasibility

The goals to construct the first large-scale corpus-based system that produces semantic representations for many possible English sentences, and to use the resulting vector-based representations to simulate human semantic behaviour are novel and bold, and success is not guaranteed. There are however, good reasons to believe that COMPOSES has the right team to pursue these goals, that the project will be at least partially successful, and that even partial success will have a high impact on the interested fields.

Strength of team: We are in an exceptionally good position to attempt the proposed integration of distributional semantics, formal semantics and experimental methods. As discussed in Sec. 1, I have a strong theoretical linguistics background as well as extensive experience in working with large corpora, distributional semantics and elicitation and analysis of linguistic data. The senior team members that will join me in the project (Raffaella Bernardi and Roberto Zamparelli) are leading experts in formal semantics, and they also have experience with corpus-based methods. The COMPOSES team will be extended with project-hired personnel to strengthen all necessary areas of expertise (including advanced machine learning methods). Finally, at the CLIC lab and at CIMeC, we have very good facilities to pursue computational and experimental work.

Work plan and delimiting the task: We articulate the project into 5 years (detailed work plan in Part B2) to make sure we have enough time to account for dependencies among the activities. We insure in

particular that there is time to extensively test the computational systems on the experimentally obtained data *after* the system has been developed and the data have been collected and analyzed. We also allocate adequate time at the end of the project to disseminate its results appropriately. We schedule periodic project reviews that will allow us to re-assess and adjust the goals of the project as we proceed. At a more technical level, we limit the compositional phenomena to a clearly spelled out subset of rules that will be extended incrementally during the project (but we will not attempt to model, for example, clause subordination or quantifier scope). We assume, moreover, that input sentences have been syntactically pre-processed, so that we only need to handle syntactic structures that match our composition grammar. Finally, we include in the experimental data sets simpler sentences and standalone phrase stimuli, to be able to evaluate the model even if we manage to implement it only partially.

Resources and preliminary evidence: COMPOSES builds on resources and tools I have already developed (a large source corpus, distributional semantic data and toolkits). In Baroni and Zamparelli (2010), we implemented a COMPOSES prototype and tested it on the task of predicting the vector of unseen adjective-noun constructions. The system worked well in general, and best with very frequent, highly polysemous and vague adjectives such as "*different"*. This suggests that scaling up at least to determiner phrases should be feasible (the step from "*different dogs"* to "*some dogs"* is not huge). Part B2 reports a pilot study showing that corpus contexts contain enough distributional cues to reliably characterize different kinds of determiners. Taken together, these preliminary results indicate that, while we cannot tell whether the method will handle all the constructions and tasks we want to model, there is no doubt that the approach will work for at least a subset of the intended constructions.

Interestingness of partial success: Even if the empirical results turned out to be disappointing, COMPOSES will still have a strong impact on many fields. First, COMPOSES should foster new approaches to the computational and empirical study of semantics above the word level. We might not be able to provide all the answers, but the questions we ask should raise a strong interest and encourage others to pursue the same goals. Second, the data sets we will collect and release will stimulate more research in the desired direction. Third, the question is not whether the model will work (it works already, at least to a certain extent, with adjective-noun combinations), but how well it will work, and for which constructions. But the latter are interesting empirical questions: we might find out, for example, that the COMPOSES approach to distributional semantics successfully models the internal structure of determiner phrases, but predication requires other methods to be developed in future research. A partial "failure" of this sort would still be an important result!

Expected impact

For decades, the study of semantics has been carried out by separate communities that communicated very little: cognitive scientists and computer scientists interested in statistical/distributional models; formal semanticists modeling logical aspects; cognitive linguists looking at semantics in a broader context, etc. By providing the first full-fledged, data-induced, general-purpose compositional semantic model of full sentences, COMPOSES should attract the interest of all these communities, and thus pave the way for other integrated approaches that will take the next big steps forward, by harnessing the strengths of different traditions. The times are ripe for such an integration, and the COMPOSES team will become an important reference point for a new community of multi-disciplinary, open-minded semantics researchers.

To distributional semantics, COMPOSES contributes a way to go beyond the word level, thus making the distributional approach amenable to a new class of semantic challenges. To compositional semantics, COMPOSES offers a path towards implementation of formal models of composition on a large scale. The vectorial representation of sentence meaning sits well with the gradient, prototype-theory-based view of language often advocated by cognitive linguists and other cognitive scientists, while making this view more explicit and amenable to computational modeling. We will not bring the state of the art in machine learning forward, but our multiple function learning scenario constitutes an excellent test bed for advanced hierarchical and meta-learning methods applied to very large data sets.

COMPOSES focuses on basic science, but there are many possible applied offshoots for a data-induced system representing sentential meaning. Applications range from question answering to query reformulation in search, to semantic technologies that could revolutionize the way we interact with computers and the Web.

Finally, while in COMPOSES we limit our study to English for feasibility reasons, the system we propose is relatively knowledge-lean: given an (automatically) annotated corpus and a composition grammar, it should be possible to create COMPOSES systems for any natural language. This, in turn, opens many new theoretical and applied avenues we hope to explore in the future.