ERC Starting Grant
Research proposal (Part B section 2 (B2))

Section 2: *The Project proposal*

a. State of the art and objectives

a.1 Objectives

Semantics is the cognitive faculty that allows us to use language to reason and communicate about states of the world and of our minds (Chierchia and McConnell-Ginet 2000). As such, it constitutes the interface of language with conceptual knowledge and other aspects of cognition (vision, decision, etc.: Jackendoff 2002). One of the longest-standing Holy Grails in linguistics, cognitive science and artificial intelligence has been to devise an artificial system endowed with human-like capabilities to understand and use natural language semantics. Computational semantic systems are of interest from multiple perspectives: for theoretical linguistics, they promise to widen the coverage of semantic analysis and free it of unrealistic idealizations (Bos 2005); for cognitive science, they can tell us which aspects of meaning can be extracted from exposure to raw data plus general learning mechanisms and which require specialized learning strategies (Landauer and Dumais 1997, Rogers and McClelland 2004); for applied research (in natural language processing, information retrieval and related areas), they can greatly improve any task where machines require semantic knowledge (question answering, information retrieval and machine translation are among the applied domains that have recently benefited from progress in data-induced semantic models; see Jurafsky and Martin 2008, Manning et al. 2008).

The desiderata for a convincing computational semantic system include: (a) 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 1994); (b) 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); (c) Agreeing with human behaviour and intuition on a variety of **semantic tasks** (if humans find "*lions have a mane"* more natural than "*lions are males",* objective statistics notwithstanding, our system should also do so). We believe moreover that points (a) to (c) should be pursued via data-intensive induction methods, to obtain a truly scalable and adaptable system.

At present, no high-coverage, general-purpose computational model exhibiting human-like behaviour on semantic tasks involving full linguistic utterances has been developed. COMPOSES intends to take a major step forward in this direction. We expect COMPOSES to be the first system that satisfies the desiderata we just listed for an explicitly delimited set of possible English constructions. Consequently, COMPOSES aims to develop a general-purpose system that

(a) learns semantic representations from naturally occurring text data on a large scale without requiring manually labeled examples;
(b) produces, compositionally, representations for larger constructions, up to full sentences;
(c) emulates human performance on many semantic tasks.

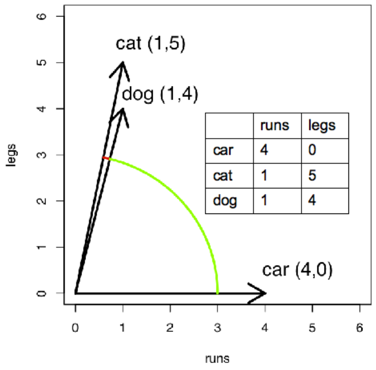
In the process of achieving these objectives, our project will produce some important intermediate results. We will develop a large **empirical base of human-elicited semantic data** to evaluate the COMPOSES method and any other computational system that produces comparable representations for full sentences. Furthermore, the results of COMPOSES (both the tools and data sets, to be publicly released, and the theoretical conclusions, to be reported in a number of articles) should **bring about a new research paradigm in semantics**, a methodology that uses up-to-date machine learning techniques to combine the highly successful approach to word meaning from the statistical/data-driven tradition with the compositional approach to utterance meaning of formal Montague-style semantics, overcoming the limits of both traditions, and bringing us closer to computational systems that fully understand and use language like humans do. The program is expected to last five years, with well defined milestones, and intermediate deliverables (detailed in the work plan below) that are independent from the full success of the final goal.

a.2 State of the art

We review here various computational and empirical semantic frameworks that are of direct relevance to COMPOSES. We will not review other important approaches to semantics, such as connectionist models (Rogers and McClelland 2004), cognitive linguistics (Croft and Cruse 2004) and the psychological and philosophical literature on concepts (Margolis and Laurence 1999), even though some of the intuitions in these studies will be represented in the project.

Distributional semantics

Considerable advances have recently been made on methods to acquire semantic representations of single words from large text collections (corpora). *Distributional semantic models* (**DSM**s, see Turney and Pantel 2010 for a survey), 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 (these raw co-occurrence counts are typically transformed with statistical weighting and dimensionality reduction techniques). Similar words tend to occur in similar contexts, and thus their distributional vectors point in similar directions: geometric distance approximates similarity in meaning. This basic idea is illustrated for an artificially small vector space in Fig. 1.



	runs	legs
car	4	0
cat	1	5
dog	1	4

Figure 1: A toy distributional semantic space.

These *data-driven* and *wide-coverage* systems – pioneering DSMs such as LSA (Landauer and Dumais 1997), HAL (Lund and Burgess 1996) and those that followed in their steps (including our own models: Baroni et al. 2010, Baroni and Lenci to appear) – have been evaluated against human behaviour and have succeeded in simulating an impressive array of lexical (i.e., word-level) semantic tasks, from detecting synonyms to concept categorization, from finding prototypical properties of concepts to predicting semantic priming. Two very attractive (and related) features of DSMs are that they are *unsupervised*, i.e., they do not require manually labeled examples of the target outputs to be trained, and they are *general-purpose*, i.e., a model is extracted once from the corpus (typically, in the form of a co-occurrence matrix), and can then be used for all sorts of different semantic tasks. Moreover, while early models (such as LSA and HAL) relied on simple contexts such as documents or words within a fixed window, more recently there has been extensive research on more linguistically sophisticated contexts that take syntactic structures or surface indicators of semantic relations into account (see our own contribution in Baroni et al. 2010, Baroni and Lenci to appear, and references there).

The great expressive power of the human semantic system certainly relies on the ability to code thousands of concepts into distinct word meanings, but we can communicate an infinity of meanings thanks to *compositionality*, the ability to combine those "atomic" word meanings into new, composite expressions. DSMs have until now made very little progress on compositionality. Some very interesting recent work (e.g., Mitchell and Lapata 2008, Erk and Padó 2008, Thater et al. 2010) has focused on how the meaning of single words is affected by the composition operation (disambiguation via co-composition: e.g., how the meaning of the verb "*run"* changes when the subject is "*water"* or "*horse",* and vice versa), rather than on the composite meaning itself. The little work done in DSMs to build vectors for composite phrases, summarized (and extended) in Mitchell and Lapata (2010), shares three highly problematic assumptions. First, compositionality is given by a single, fixed mathematical operation (such as sum, weighted sum, dimension-wise product) on the vectors representing the words to be composed. Second, the composite phrases considered are invariably made of two *content* words (words with rich lexical content, such as nouns, verbs and adjectives). *Grammatical words* (e.g., articles, prepositions, quantifiers, or modal verbs), that provide the structural scaffolding of all but the most elementary sentences, are completely missing. Third, the syntax is oversimplified (and typically very "flat", see e.g., Coecke et al. to appear).

Our goal is to maintain the advantages of the DSMs research line, which is suitable for achieving our objective (a), and extend it to phrases and full sentences, overcoming the limits of the state-of-the-art systems.

Formal semantics

Formal semantics (**FS**) (Chierchia and McConnell-Ginet 2000), the model theoretical research program harking back to the seminal work by Montague (1970), has in many respects opposite strengths and weakness with respect to distributional semantics: it has produced very sophisticated models of meaning composition (our objective (b)), but it has not implemented them in data-driven models with large coverage of word meanings. More specifically, FS has furthered our understanding of the *functional* scaffolding that keeps sentences together, but it has largely ignored (with important exceptions, such as Pustejovsky 1995 or Asher to appear) the problem of capturing the meaning of single *content* words and understanding how this meaning is dynamically specified in the composition process (in most of the literature, it is tacitly assumed that when two semantically ambiguous words are combined in the syntax, they are somehow disambiguated before being semantically combined).

Montague and most of his followers were not interested in cognitive aspects, but in an elegant and simple mathematical framework for natural language. As a consequence, their ideas and intuitions have only recently been checked for cognitive plausibility and tested against data that go beyond paper-and-pencil case studies. Yet, even Bos (2005) - a successful wide-coverage system able to compute first order logic representations for English sentences - suffers for the lack of a proper lexical semantics. Some interesting recent attempts to combine formal and distributional semantics have come from the mathematical/logical modeling tradition (Clark and Pulman 2007, Clarke 2007, Coecke et al. to appear). These proposals have focused on formalizing integrated models, and have not come up, yet, with empirically validated ideas about how such models can be instantiated, on a large scale, from corpus data. We will follow developments in this area with great attention.

The key ideas from the FS framework that we intend to operationalize are: (i) the principle of compositionality, together with the assumption that *the construction of meaning is guided by syntactic structure*; (ii) the view of words (and phrases) as *complete* (e.g., proper names) or *incomplete* (e.g., attributive adjectives) expressions, and consequently the representation of the latter as *functions*; (iii) the method of working with *fragments* of natural language in order to isolate the phenomena under investigation.

An important point concerning the relation between FS and our research program must be clarified. We aim to design a large-coverage semantic system built on vectorial lexical representations. Insofar as our system is to be tested on the ability to draw inferences, it must deal with the notion of *logical consequence*, which is one of the centerpieces of FS. This might seem to be at odds with the fact that all the knowledge in our system comes from corpora, which are blatantly inconsistent, if only for the massive referential ambiguity of the pronouns and descriptions they contain ("*he is tall"* and "*he is short"* will be amply represented in any corpus). How can we hope that our representations for sentences will give us correct, FS-style inferences? The answer is that we believe that, while corpora do not easily yield *episodic* knowledge (although there is an important tradition of corpus-based information extraction; see Cowie and Wilks 2000), they are extremely rich in *generic knowledge*: the information that "*companies have employees*", that "*ostriches are birds*" and yet "*do not fly*", plus millions of other commonsense notions will be represented in a large Enough corpus (Herdağdelen and Baroni to appear). Our goal is to design our vector representations for sentences is such a way that they can pick up these regularities and use them to draw *generic inferences*. Thus, our approach to truth is complementary to that of FS, which has not dealt with habitual/generic truths until relatively recently, and in which the issue of which sort of conclusions can be drawn from generic statements is quite debated (see Cohen 1999, ch.7). Indeed, one of the hardest problems this project can address, and one which we do not yet know how to solve, is how to integrate new episodic information from a specific world model (e.g., "*This dog doesn't bark"*) with the generic background of knowledge obtained from the corpora (*"Dogs bark"*). Capturing logical entailment is, however, only one aspect of our enterprise. Overall, we believe that its success can be more meaningfully measured by the system ability to scale to novel constructions and by the totality of cognitively plausible semantic tasks that it can perform.

Machine learning

The last 20 years have seen the establishment of *supervised machine learning* as a very successful technique to learn arbitrary functions from labeled input/output 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). These methods have been extensively applied to various aspects of semantics, such as semantic role labeling, where the arguments of a verb are automatically assigned a role such as agent, instrument, etc. (Jurafsky and Martin 2008, ch. 19 and 20). The machine learning approach to semantics is often empirically very successful, but it has two major drawbacks: (i) to work well, it needs plenty of manually labeled data, and (ii) these data – often together with the way in which they are represented – change from task to task, so that it is not clear how to develop a general-purpose semantic system using the method. Machine learning systems are typically optimized for a single task, or a small set of related tasks.

Empirical base

To meet the desideratum that a computational system should show human-like behaviour on tasks tapping into semantic knowledge (objective (c)), we must design *semantic tasks* that allow such comparison, and collect data about human performance on such tasks. We will focus on *similarity, plausibility* and *inferencing.* First, given that COMPOSES represents sentence meaning in terms of vectors, it is natural to ask whether the distance between such vectors reflect intuitions about the *semantic similarity* of the corresponding sentences. An extensive literature in cognitive science, information retrieval and computational linguistics has shown that subjects produce reliable semantic similarity judgments about pairs of words/concepts, and has defended the theoretical interest of the very notion of semantic similarity (see Medin et al. 1993, McDonald 2000 and references there). It is not obvious how to evaluate the similarity of sentences: sentences that talk about *similar topics*, making *similar claims* about them, with *parallel structures* and *similar perlocutionary effects* will most likely be judged similar, but probably a subset of these criteria will suffice. Thus, every comparison in this domain will run along multiple dimensions, and it is expected to generate interesting data not only about the model, but about how humans judge complex similarities.

Second, we would like our model to provide human-like assessment of how *semantically plausible* a sentence is (a reflex of generic knowledge). In collecting such judgments, we will take inspiration from experimental linguistics. In an attempt to make the traditional grammaticality judgment methodology more empirically sound, some linguists have recently developed experimental methods to elicit graded *plausibility judgments* about sentences, such as pseudo-continuous Likert scales, magnitude estimation and speeded judgments (e.g., Weskott and Fanselow 2008, Bader and Häussler 2010). The reliability of linguistic judgments of various kinds has also received attention in computational linguistics, where sophisticated methods have been proposed to quantify the degree of agreement of different judges (Artstein and Poesio 2008) – again, we will take these techniques into account.

For all rating tasks, we would like to access a large pool of English speakers. We will exploit Web surveying techniques, as we have successfully done in the past (Baroni, Guevara and Zamparelli 2009). A recent and exciting development is the possibility of using *crowdsourcing services,* that give access to a large community of subjects that perform simple tasks for a small accomplishment-based fee. Recent extensive tests (summarized in Munro et al. 2010) indicate that linguistic experiments conducted in this way produce very reliable results, and similarity/plausibility judgments are a very well-suited task for this data collection method, that we will thus adopt in COMPOSES.

Another perspective on evaluation, closer to the traditional concerns of formal semantics, pertains to recognizing certain logical relations among sentences: a system "understanding" semantics should be able, for example, to tell that *"Penguins do not fly"* entails that *"Some birds do not hover"*. In computational linguistics there has recently been wide interest in the *recognizing textual entailment* task (Dagan et al. 2009), whereby systems must decide whether an utterance or paragraph entails another. Following this tradition, but focusing more specifically on the strictly semantic aspects of entailment (MacCartney and Manning 2009), we will develop data sets illustrating various entailment relations that we want our models to capture.

a.3 Progress beyond the state of the art

We aim to achieve a data-driven computational semantics system that, starting from a properly rich lexical semantic representation, builds a meaning representation for phrases and sentences compositionally and in accordance with the behaviour of human beings in a variety of well defined semantic tasks. We believe our aim can be achieved by a novel integration of the strengths of the DSM and FS approaches, and that the crucial step is the ability to build/test *distributional semantic representations* (henceforth **DSR**s) for *phrases* (e.g., "*red dog"*), rather than words. In Baroni and Zamparelli (2010), we have shown that these vectors can be obtained using a simple type system and machine learning techniques that induce the actual compositional functions from corpora. Here we want to extend and generalize this method.

From formal semantics, COMPOSES borrows the syntax-guided calculus of composition, as well as the idea of complete and incomplete expressions; from DSM, the idea that an important component of the meaning of all linguistic expressions is encoded in their distributional vectors. This view of meaning is orthogonal to the denotational approach to meaning pursued in FS. While it should be quite possible to give a denotational interpretation for our phrasal vectors (though this is not among the goals of the project), they can in fact be more properly seen as a computational way to tackle the Fregean *sense/reference* distinction: two expressions with the same referent (say, "*the morning star"* and "*the evening star"*) will have distinct vectors, corresponding to their different "senses" (operationalized in patterns of usage).

In COMPOSES, we abandon the DSM assumption that words are uniformly represented by distributional vectors, to propose that each lexical category has potentially two **distributional types**: the type of *complete expressions*, that are indeed represented by distributional vectors, and the type of *incomplete expressions*, that are represented by *distributional functions* mapping vectors onto vectors (or, in some cases, from/onto other mathematical structures). Some categories use only one type (proper names are always complete/vectors; determiners are always incomplete expressions which take nouns as their arguments); others can use both types, in particular nouns (in their relational/non-relational uses: "*John is a father/a father of twins")*, and adjectives, which could be vectors in predicative position and functions in attributive position.

We do not abandon the **unsupervised and data-driven approach** of DSMs, proposing that both the vectors and the functions from vector to vector can be learned from the corpus, by directly extracting examples of the functions' input and output objects, and using them to optimize the mapping they perform. While we model a relatively limited number of composition processes via our formal composition grammar, we can harness the full power of data-driven DSMs to harvest thousands of word representations (corpus-extracted vectors and functions) to instantiate the composition rules, thus meeting our large-coverage objective. The most serious problem faced by any attempt to give a vectorial representation of grammatical words is that, since such words are so widespread, their context becomes semantically uninformative. Our approach (first illustrated for adjective-noun constructions in Baroni and Zamparelli 2010) solves this problem by treating all grammatical words as functions. For example, to learn the "*some"* function from an noun *N* vector onto the corresponding "*some N"* vector, we obtain training data by scanning the corpus for occurrences of each *N* and the corresponding "*some N"* phrases, we construct vectors from their corpus occurrences, then use the resulting <*N, "some N"*> vector pairs as if they were manually labeled training data to be fed into a **supervised regression algorithm.** This approach, in which the algorithm creates its own training data, can be seen as part of the current efforts to harness the power of supervised learning while avoiding or minimizing annotation cost (Abney 2007). The problem of lack of generality affecting machine learning methods is addressed by using machine learning only as an intermediate layer to construct functions that are then applied in the general grammar-driven calculus of composition to generate sentences.

Finally, on the **evaluation** side, our main novel contribution will be the collection and analysis of new subject-elicited data pertaining to semantic similarity and plausibility at the sentential level.

a.4 Expected impact

By providing the first full-fledged, data-induced, general-purpose compositional semantic model of full sentences, COMPOSES should attract the interest of different research communities interested in semantics. 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. Moreover, at a lexical level, the system will allow us to build a new generation of DSMs for words that uses as vector dimensions not only the words nearby, but also the representations for neighbouring phrasal nodes at various levels of generality (see WP2.3). Our hypothesis is that these word vectors could more accurately capture subtle subcategorization and collocation effects. The word vectors so obtained could then be fed back into the system, bootstrapping more accurate DSRs for phrases and sentences.

To compositional semantics and theoretical linguistics in general, COMPOSES offers a path towards implementation of formal models of composition on a large scale. We should be able moreover to characterize, in terms of DSMs, a host of notoriously elusive semantic relations. For instance, the vector for *"clever dog"* could be compared to those for "*intelligent dog"* (synonym), *"clever animal"* (upward entailment), *"clever bobtail"* (downward entailment), *"stuffed dog"* (non-subsective), *"Italian clever dog"* (wrong order)*,* etc. Our hypothesis is that these relations can be recognized from some mathematical characteristic of their vectors, and that a system capable of categorizing them could generalize to other modifiers. From a discourse standpoint, a DSM model for sentences can be used to study a variety of discourse relations (e.g., paraphrases, contradiction, various types of similarity, generic inference, elaboration, etc.).

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 than in most current cognitive linguistics work.

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. 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. Multilingualism, in turn, opens many new theoretical and applied research avenues (e.g., vector-based comparisons of translations of the same text).

b. Methodology

b.1 Work plan

Timeline

	M6	M12	M18	M24	M30	M36	M42	M48	M54	M60
WP1: Data, tools and infrastructure	■	■	■	■	■	■	■	■	■	■
WP1.1: Computational infrastructure	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
WP1.2: Data and tool packaging and release						▨	▨	▨	▨	▨
WP2: Computational modeling	■	■	■	■	■	■	■	■		
WP2.1: Composition grammar	▨	▨	▨	▨	▨	▨	▨			
WP2.2: Distributional space construction	▨	▨	▨	▨	▨	▨				
WP2.3: Learning distributional functions		▨	▨	▨	▨	▨	▨	▨		
WP3: Semantic tasks for empirical evaluation	■	■	■	■	■	■	■	■	■	
WP3.1: Plausibility ratings	▨	▨	▨	▨	▨	▨	▨			
WP3.2: Semantic similarity ratings	▨	▨	▨	▨	▨	▨	▨			
WP3.3: Entailment	▨	▨	▨	▨						
WP3.4: Model validation		▨	▨	▨	▨	▨	▨	▨	▨	
WP4: Coordination and dissemination	■	■	■	■	■	■	■	■	■	■
WP4.1: Project management	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
WP4.2: Dissemination	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨

Work Package 1: Data, tools and infrastructure [Months 1-60]
Objectives: set up and maintain computational infrastructure; package and document data and tools to be released.

WP1 Activity 1: Computational infrastructure (Lead: PI; participants: post-doc 1) [Months 1-60]
In order to run the intensive computations required by the project (Activities WP2.2, WP2.3 and WP3.4), we will use COMPOSES funds to extend our current server cluster setup to (minimally) 10 16GB RAM blades and 20TB high quality (NetApp) storage space. On top of this hardware layer, we will implement a MapReduce/Hadoop architecture for transparent parallelization of common operations such as feature extraction, counting and cross-validated parameter estimation . Many tasks we need to implement involve multiple application of simple and independent computations to very large data sets, and are thus well-suited to the MapReduce paradigm. Activity 1 lasts for the whole duration of the project (beyond its last milestone) to account for any unforeseen infrastructure roadblock that might arise at any time.

WP1 Activity 2: Data and tool packaging and release (Lead: PI; participants: everybody) [Months 34-60]
Special care will be taken in packaging the software tools and resources produced by the project, so that they will be of use to various communities (linguists, language engineers, cognitive scientists...), furthering the impact of the project by making it easier, on the one hand, for other groups to get started in the COMPOSES research program, and, on the other, by providing semantic resources that could be used for completely different purposes. We have extensive experience with resource construction and distribution, and we have already developed tools for creating and manipulating distributional semantic models, including some that we made publicly available. Within COMPOSES, we will extend these tools, as well as improving their integration and documentation. In terms of data, we will document and make available both pre-trained distributional models and all our evaluation data sets. The latter will include sentence lists, their properties and relations, subject ratings, and general descriptive statistics that might be of use to other researchers (e.g., frequency of occurrence of the words and phrases in the sentences).

Milestones and deliverables

M1.1.1	Extended server up and running	M6
M1.1.2	Hadoop/MapReduce architecture ready	M12
M1.1.3-5	Annual infrastructure review	M24,36,48
M1.2.1	Precompiled semantic space resources packaged	M39
M1.2.2	Experimental data set streamlining	M48
M1.2.3	Code cleanup and review	M57
D1.1.1	Architecture documentation	M13
D1.2.1	Public precompiled semantic spaces release	M40
D1.2.2	Public experimental data set release	M50

D1.2.3	Public code release	M59

Work Package 2: Computational modeling [Months 1-48]
Objective: implementation of semantic representation and computation system.

Given an input sentence (we assume syntactic pre-processing to transform it into the structure required by the COMPOSES grammar: integration with syntactic parsing is not an objective of the project), the COMPOSES system must first retrieve distributional vectors and functions corresponding to all constituents in the sentence, and then apply them in the order determined by the grammar, to compute the output vector representing the sentence (please refer back to the schematic illustration of the process in Fig. 1 of Part B1). To perform these steps, the system must have access to a grammar spelling out the calculus of composition (Activity 1), precompiled distributional vectors extracted from the corpus (Activity 2) and distributional functions trained on corpus data (Activity 3).

WP2 Activity 1: Composition grammar (Lead: Bernardi; participants: PI, Zamparelli, PhD student 2) [Months 1-39]
Following the methodology of Montague (1970), we will model only a small subset of the possible composition rules needed by a full grammar of English. Still, form this small set we can generate infinitely many sentences made of the large set of words from our corpus-derived vocabulary of DSRs. There are many important questions that we will *not* tackle within the limits of COMPOSES, such as how to handle sentential subordination or quantifier scope (though we can approach Wh-movement; see WP2.2). We start with something limited, but with clear ideas about how to make it concrete in a large-scale distributional semantic implementation. Further complexities can be added later.

We propose two versions of the COMPOSES grammar fragment: G1.0 and G2.0. The primary goal of G1.0 is to reach the sentential level, using a few “safe” rules that build on previous work of ours. This will allow us to test the robustness of sentence-level DSRs, compare various methods to obtain word-level DSRs and different learning functions. With this methodology in place, we will try G2.0, which is linguistically more motivated and more general, and which should allow more general testing of inferences. In G1.0, nouns (**N**: “*dog”*) are uniformly treated as complete expressions (thus vectors, not functions). Attributive adjectives (**Adj**: “*red”* in “*red dog”*) map nouns onto nouns. Determiners (**Det**: “*the”, “a”*, “*any”*, “*some”*, “*many”,* “*every”*) map nouns onto noun phrases (**NP**; DP in much current syntax; we ignore the distinction), which are complete expressions like Ns (vectors, not functions). In G1.0, we do not treat quantifiers as diadic operators, as in FS; analyses more in line with FS (in which a quantifier takes a noun and returns a function) will be explored for G2.0. Prepositions (**P**: “*of”*, “*in”*, “*on”*, “*under”*, “*over”*, “*from”, “with”*, “*without”*, …) are functions mapping NPs onto preposition phrases (PPs). Prepositional phrases (**PP**: “*of the dog”*) are functions from Ns onto (complex) Ns (from “*tail”* to “*tail of the dog”*). NP conjunction takes two NPs and returns a single NP (from “*the dog”* and “*the cat”* to “*the dog and the cat”*). Predicative adjectives *(“is red”* in “*The house is red”;* in G1.0 we ignore the copula) are functions that take NPs and return sentences (complete expressions). The G1.0 rules represent a number of composition types, including functions taking one (e.g., attributive adjectives) or more (conjunction) complete expressions and returning other complete expressions, or (in the case of prepositions) returning functions. With G1.0, we can already generate rather complex sentences: “*The sad red dog with a hairy tail and some stray cats from the neighbourhood are dangerous”* and test their semantic properties and discourse relations.

G2.0 adds to G1.0 the following elements: **names** of peoples and places (obtained from corpus pre-processing); **relational Ns** (“*mother of”, “end of”*), modeled as functions from NPs to Ns (from *“the semester”* to *“end of the semester”*); **other Dets**: *“which”, “how many”, “all”,* and bare plurals; like in G1.0, they map Ns onto Ns. The last two appear in generic sentences. Bare plurals exemplify a purely position-driven composition process, while Wh-elements will allow us to test interrogative sentences. In G2.0, unlike in G1.0, we will treat predicative Adjs as complete elements, i.e vectors (as in Guevara 2010), leaving the composition with their subjects to the **copula,** a function from NP (the subject) and Adj/NPd (the predicate) to sentences. Finally, G2.0 will contain **predicate modifiers** (*“very”* in *“very red dog”*), functions from predicative Adjs (vectors) to other predicative Adjs, and **predicate conjunction** via *“and”,* a function from pairs of Adjs/NP to a single Adj/NP (e.g., from *“heavy”* and *“large”* to *“heavy and large”*. This is usually the intersective *“and”*). Further natural extensions (G3.0) include **intransitive verbs** (treated as vectors, like predicative Adjs), combined with subjects via **T**(ense) (a position-driven function from NPs and V to sentences, similar to the copula) and **predicate negation** (“*not”,* from predicates to predicates).

Activity 1 will keep refining and extending these grammars in an open feedback loop with Activity 3 and empirical tests.

WP2 Activity 2: Distributional space construction (Lead: PI; participants: senior staff, PhD students, post-doc 1) [Months 1-36]

The DSRs for single words and larger constituents (e.g., adjective-noun or determiner noun constructions) to be used as input and output examples in distributional function training are induced from patterns of occurrence of the relevant expressions in the source corpus. Activity 2 focuses on optimizing the distributional representation of linguistic expressions in terms of corpus-extracted features. We will use the very large and linguistically processed corpus described in Baroni and Lenci (to appear) (about 3 billion words). By exploiting the linguistic annotation already present in the corpus (including lemmatization, part-of-speech tags and syntactic dependency structures), we can extract features recording properties of the context at different levels of linguistic sophistication: inflected or lemmatized collocates in a fixed window, syntactically-mediated collocates ("*subj-kill"* and "*obj-kill"* as opposed to simply "*kill"*), morphological features (number, inflection), syntactic templates ("*subj-verb-prep"*), and even (with further corpus processing) shallow cues of discourse structure. Dislocated constituents can be modeled by copying ("reconstructing") a Wh-phrase in its base position (e.g., turning "*which boy did Mary see?"* into *"which boy did Mary see which boy"*). Other forms of corpus pre-processing, such as *named entity recognition* and possibly *anaphora resolution* might be applied to handle proper names and sentences with pronouns. To keep the function learning problem feasible (while possibly also improving performance by lowering the noise), we will apply feature selection and/or dimensionality reduction techniques to the full co-occurrence matrix.

WP2 Activity 3: Learning distributional functions (Lead: PI; participants: Zamparelli, post-doc 1, PhD student 1) [Months 7-48]

For each of the composition operations in our Activity 1 grammars, we need to estimate a distributional function mapping the DSR of the input expression(s) to that of the output expression. Baroni and Zamparelli (2010) treat attributive adjective functions as linear maps from and to *n*-dimensional noun vectors (e.g., from "*book"* to "*red book"*), encoded in *nxn* square matrices (the output vector is obtained by multiplying the matrix by the input vector). The weights in the matrix representing an adjective are estimated using partial least-squares regression (Hastie et al. 2009, 3.4) trained on corpus-derived input-output example pairs (e.g., corpus-derived vectors for < "*book", "red book"*>, < "*army", "red army"*>, etc.). Although we might look into different estimation methods, we will at least initially focus on the same approach, i.e., we will assume that distributional functions are linear and the learning problem one of matrix estimation. The Baroni and Zamparelli method can applied "as is" to other functions from vectors onto vectors. For example, the "*some"* matrix can be learned from pairs such as <"*dog", "some dog"*>. The approach should also work for functions from vector pairs onto vectors: *NP and NP* can be represented by a *nx2n* matrix, the input for the mapping is the concatenation of the vectors representing the two NPs, and the matrix is estimated from example inputs like "*cats"* and "*dogs"* with corresponding output "*cats and dogs"*. In a more explorative way, we will apply a similar approach to learning vector-to-function functions such as prepositions (encoded in n^2xn matrices, where the output n^2 dimensions constitute the preposition matrix weights).

Clearly, each individual sentence or complex phrase will be rare, if present at all, even in a very large corpus; thus, their DSR will be sparse, or impossible to build (think of searching for the corpus collocates of "*the old pink car is very heavy"*). One research topic crucial for our project will thus be how to overcome this data sparseness problem. We plan to tackle the problem in two ways. First, we never use cases as complex as "*the old pink car is very heavy"* for training. Rather, we train the "*be"* function (in G2.0) on "*the car is heavy",* and thousand other simple cases which *are* represented in the corpus, then use the trained function to generate the complex cases. This is part of why it is is so important to have human- and inference-based methods to evaluate the quality of the output, as described in WP3: in most cases, there is no way to test that the model predictions correspond to *attested* meanings/contexts, as we could do with *"red dog"* (this is, after all, the consequence of having a true *generative* system). Second, we can make the input and output vectors denser by soft or hard vector clustering techniques (Hastie et al. 2009, ch. 14), whose net effect is that each constituent can be associated not with a single DSR, but with a set of increasingly general DSRs. Greatly simplifying, given the string *"The hunter shot. Pink flamingos flew away",* the DSR for *"the hunter shot"* could be constructed on the basis of the following adjacent context: {*"pink", "flamingos", "[pink flamingos]", "[COLORED BIRDS]", "[ANIMAL]", ...*}, where the uppercased categories are superordinates of the NP "*pink flamingos"* derived via clustering. The rationale is that even if "*the hunter shot"* appears only once with *"pink flamingos",* it will appear many times with ANIMAL.

Another important issue pertains to capturing *similarity across functions*: we should not learn predicative *"red"* in a vacuum, but exploit its similarity to the attributive "*red"* function and to the predicative "*yellow"* vector (or function, in G1.0). Multi-level machine learning methods such as hierarchical regression, Markov Logic Networks, Deep Learning and, more in general, graphical models (Koller and Friedman 2009) should come handy to tackle this sort of generalization problem.

Milestones and deliverables

M2.1.1-3	Grammar specifications completed	M6,27,39
M2.2.1-3	Extraction of corpus-based semantic spaces done	M8,22,34
M2.3.1	Re-implementation of composition models from earlier literature done	M12
M2.3.2	Implementation of first version of COMPOSES distributional functions done	M15
M2.3.3-5	Advanced implementations of distributional functions done	M25,35,45
D2.1.1-3	Reports about grammar specification	M6,27,39
D2.2.1	Internal release of trained semantic space resource	M36
D2.3.1	Internal release of COMPOSES modeling toolkit	M48

Work Package 3: Semantic tasks for empirical evaluation [Months 1-50]

Objectives: collection of human-elicited and expert-constructed data sets about various aspects of sentential meaning; evaluation of computational systems on semantic tasks.

Given the innovative representation of sentences as vectors, we must develop new *semantic tasks* that illustrate the meaningfulness of this representation and assess the quality of the COMPOSES models (and, when it is possible to adapt them to the tasks, alternative models from the earlier literature). An important by-product of our experiments will be the construction of empirical human-elicited semantic data sets that we will share with the research community. We design various data sets tapping into different aspects of semantics, while making sure that the sentences they contain are covered by the generation system detailed in WP2, both in terms of words and constructions.

We elicit subject judgments about how plausible sentences are (Activity 1), and about sentence similarity (Activity 2). Pilot studies will be run by recruiting native and advanced non-native English speakers from the University of Trento large international community. A much wider subject pool will then be reached by Web surveying techniques and crowdsourcing. For the next data set, we do not rely on subject ratings. We apply instead various linguistic criteria to build a database of sentences entertaining various logical entailment relations (Activity 3). By controlling different factors in data set creation, the three tasks we use (plausibility, similarity and entailment) allow us to check for how both humans and the computational systems respond to a wide number of semantic phenomena. Activity 4, finally, implements the actual evaluation of the computational models, in particular by measuring their performance on test data collected in the previous activities.

WP3 Activity 1: Plausibility ratings (Lead: PI; participants: Zamparelli, post-doc 2, PhD student 2) [Months 1-40]

In a first batch of experiments, by adopting elicitation techniques from experimental linguistics, we will collect subjects' ratings about the plausibility of various sentences. We will first run a series of pilot studies to choose the most reliable rating method among Likert scales, magnitude estimation, sliding bars, etc. We will also explore various data collection settings, aiming in particular at large-scale surveys via crowdsourcing services. After converging on the optimal data collection strategy, we will use it to collect plausibility ratings about sentences selected to tap into various potentially interesting factors affecting plausibility, such as genericity effects (we would expect higher ratings for "*some dogs are crippled"* than for "*all dogs are crippled"*), contradiction ("*mortal beings are immortal"*) and conceptual anomaly ("*talkative tables are eatable"*). Here and below, sentences will be controlled for potential confounds such as the frequency of occurrence of the sentence as a whole and its constituents.

WP3 Activity 2: Semantic similarity ratings (Lead: PI; participants: Bernardi, post-doc 2, PhD student 2) [Months 1-40]

If sentential meaning is, at some level, represented by a vector, sentential semantic similarity must be a meaningful notion, since we can meaningfully talk about how similar a pair of vectors are (as in Fig. 1 above). To verify this general hypothesis, and to see whether the COMPOSES-generated vectors in particular are good similarity predictors, we will prepare a data set based on subject-elicited semantic similarity judgments for whole sentences. There is a large literature on how various sentential contexts affect similarity judgments and other behavioural tasks (e.g., McDonald 2000, Mitchell and Lapata 2008). Mitchell and Lapata (2010) have shown, moreover, that people can produce reasonably coherent judgments about the similarity of phrases (see also Hampton's work on prototypically judgments about phrases, e.g., Hampton 1991). Ours is however, to the best of our knowledge, the first attempt to collect similarity judgments for full sentences with different structures. Thus, we must first of all establish that we can indeed elicit coherent judgments. This will be an interesting result by itself, providing evidence for our sentence-meaning-as-vector view.

After solving the same methodological issues as for Activity 1 (rating method, subject pool, etc.), we will use the paradigm to investigate questions such as: How does inserting/removing modifiers affect similarity between a pair of sentences? How about changing a quantifier? What is the effect of replacing lexical items with similar words while keeping the functional structure intact? How does the presence of entailment or contradiction affect the judgment? Are there interactions among these and other factors? Sentence similarity judgments will also be used (similarly to Mitchell and Lapata 2008) to elicit data related to the disambiguation of polysemous words in context (see discussion in Sec. a). Subjects will probably find "*The cucumber is old"* more similar to: "*Food is rotten"* than to "*Food is ancient"* – will our models capture this intuition?

WP3 Activity 3: Entailment (Lead: Zamparelli; participants: PI, Bernardi, PhD student 2) [Months 1-24]

Our final evaluation strand pertains to capturing entailment relations between sentences. Predicting entailment relations between propositions is a methodological pillar of formal semantics, and recognizing textual entailment is also a popular task in natural language processing. Unfortunately, we cannot rely on existing textual entailment data sets, since they require handling too many modules that are peripheral to our goals (identifying company names, syntactic parsing, etc.). We will instead take inspiration from work on constructing test cases that highlight the strictly semantic properties of entailment, such as down- or upward-entailment effects of quantifiers on implication (MacCartney and Manning, 2009). We will make a list of properties that affect entailment and that are encountered in the kinds of sentences that our models generate, such as the direction of implication of quantified NPs ("*all dogs"* implies "*some dogs"*) or lexical inclusion ("*dogs"* implies "*animal"*), and we will then use this list to systematically generate sentence pairs in various entailment relations (bidirectional: "*some canine is brown"/"a dog is brown";* one-way: "*all dogs are brown"*/"*some animals are coloured";* no entailment: "*some dog is immortal"*/"*many animals are mortal";* contradiction: "*some dog is immortal"/"all animals are dead"*). We will not ask subjects to rate the sentences, but we will decide the relation they stand in based on standard linguistic and logical tests.

WP3 Activity 4: Model validation (Lead: PI; Participants: everybody) [Months 7-50]

During development, the quality of our semantic space models will be assessed by standard lexical semantic tasks, such as the ones we collected in Baroni and Lenci (to appear). Distributional functions will be evaluated (as in Baroni and Zamparelli 2010) in terms of approximation of function output on held-out data. The really crucial evidence will come however from letting the models face the semantic tasks described in the previous activities, and compare their responses to the ones given by humans.

For sentence similarity, we can follow the standard approach in distributional semantics (Turney and Pantel 2010, see also Fig. 1 above) of computing the cosine of the angle formed by two sentence-representing vectors and correlating this cosine with average similarity scores for the same sentence pairs (the higher the cosine, the nearer the vectors, the more similar we predict sentences to be). For plausibility and entailment, we will pursue a supervised approach in which, given manually labeled/scored examples of sentences or sentence pairs, we test whether a learning algorithm trained on the COMPOSES-produced vector representations of such sentences can guess the label/score of unseen sentences or sentence pairs (note that the manual annotation here is to adapt COMPOSES to the tasks, and not part of its sentence generation system). For example, we can train a supervised algorithm on a set of COMPOSES vectors representing conceptually well-formed ("*spotty dogs are cute"*) and anomalous ("*brown episodes are ugly")* sentences, with the scores assigned by humans to these sentences as the target of learning, and then let the trained system provide scores for more or less anomalous sentences that were not seen in training. An important part of the evaluation, beyond performance analysis, will consist in the post-hoc evaluation of the features that have the most significant impact on performance, since they should be very informative about how certain semantic properties of sentences manifest themselves in distributional semantic space. We could find out, for example, that the vectors of anomalous sentences tend to be shorter, or have low values in certain feature subspaces (subspace clustering techniques might come handy here), or that asymmetric entailment tends to correlate with certain geometric patterns in semantic space, along the lines of recent work on lexical inference in vector spaces (e.g., Erk 2009).

Milestones and deliverables

M3.1-3.1	General data collection strategy design sketched	M6
M3.1-3.2	Stimulus set design ready	M14

M3.1,2.3	Pilot elicitation experiments conducted	M24
M3.1,2.4	Large-scale elicitation experiments finished	M36
M3.1,2.5	Data analysis finished	M40
M3.3.1-3	Evaluation of semantic spaces on standard lexical tasks done	M9,23,35
M3.3.4-7	Evaluation of distributional functions by approximation of output vectors done	M16,26,36,46
M3.3.8,9	First and second round evaluation of COMPOSES and alternative models on entailment data completed	M28,48
M3.3.10-14	First and second round evaluations of COMPOSES and alternative models on plausibility and similarity data completed	M42,48
D3.1-3.1,2	Stimulus sets: initial, revised	M14,24
D3.1-2.3	Internal experiment result data set release: raw, processed	M37,40
D3.3.1-3	Evaluation reports	M28,42,50

Work Package 4: Coordination and dissemination [Months 1-60]

Objectives: oversee project-internal communication flow; assess progress and adjust targets; oversee dissemination by publications and other means.

WP4 Activity 1: Project management (Lead: PI; participants: everybody) [Months 1-60]

The PI and the two senior team members will guarantee that the project is on track and communication flows. Meetings of the project personnel will be held on demand, but minimally on a bi-weekly basis. The 3 external consultants will be invited after 18, 36 and 54 months to assess the progress of the project, identify open issues, suggest priorities, as well as consider possible avenues for post-COMPOSES projects. The second review, in particular, will thoroughly consider the status of the project shortly after its middle point, and re-adjust its goals accordingly.

WP4 Activity 2: Dissemination (Lead: PI; participants: everybody) [Months 1-60]

The PI and team members will insure that enough project time is dedicated to the dissemination of the project results. COMPOSES will constitute the central activity of the CLIC lab for the 5 years of its duration. Thus, the team will insure a strong link between COMPOSES and CLIC-wide initiatives, in particular by organizing semester-long thematic seminars on themes of general CLIC interest but relevant to COMPOSES. From the start, we will post a simple but informative COMPOSES website, that will eventually collect all data sets, tools and publications produced by the project. Proceeding publications will be the main instrument for periodic updates about the project (candidate venues include ACL, IWCS and CogSci), but our main publication target will be to produce a series of journal articles addressing various research communities. We will publish at least one article in a theoretical linguistics/semantics journal (about the theoretical aspects of our approach to compositionality), one aimed at computational linguistics/NLP (focusing on computational and machine learning aspects and performance), one in a general cognitive science journal (mostly about the empirical evidence), one in a journal focusing on linguistic/cognitive resources/tools, and possibly an overall summary to be published in a huge-impact general science journal. In the last year of the project, coinciding with the last external consultant review, we will organize a workshop about the topics of the project, inviting the consultants as well as other personalities in the relevant fields.

Milestones and deliverables

M4.1.1	External personnel (excluding students) hired	M6
M4.1.2	Server extension purchased	M2
M4.1.3,4	Preliminary and final opinion of UNITN Ethical committee obtained	M6,16
M4.1.5,6	PhD students selected	M12,24
M4.1.7-9	External consultant reviews	M18,36,54
M4.2.1,2	Project website up (M4.2.1); all project deliverables available from website (M4.2.2)	M3,60
M4.2.3,4	Organization of CLIC thematic seminars on COMPOSES topics	M18,42
M4.2.5	Final workshop	M54
D4.1.1-3	Review reports	M19,37,55
D4.1.4,5	PhD theses defended	M48,60
D4.2.1-4	4 conference papers: 3 on computational modeling, 1 on elicitation experiments (milestones on submission dates)	M28,40,42,48
D4.2.5-9	5 journal articles: about elicitation experiments, computational modeling, compositionality, tools and resources, general scientific significance (dates submitted)	M44,50,54,60,60

Dependencies and risk management

The detailed 5-year work plan above insures an optimal relative timing of the activities. It allocates ample time, in particular, to model evaluation *after* the models have been implemented and the benchmark data sets have been constructed, as well as to final dissemination of the project results. The regular meetings of the COMPOSES team and the periodic consultant reviews will assess progress, and re-adjust the objectives accordingly.

We briefly comment here on two critical dependencies and the connected risks. (*1) WP2 modeling undershoots its target, failing to implement all the rules we would like to account for.* We will make sure that the WP3 data sets include sentences of different degrees of complexity, as well as phrases (e.g., adjective-noun constructions) as standalone stimuli, to allow incremental evaluation of the WP2 models and meaningful evaluation even if it turns out that some of the more advanced aspects of the WP2 grammars are not feasible to implement. In any case, the full WP3 subject-elicited data will be released, and they will constitute a precious resource for future empirical semantic studies. The partial success and specific failures of WP2 modeling will also be documented, and they will be of interest by themselves, and in paving the way for future work. *(2) Some WP3 experiments do not provide meaningful results* (e.g., subjects do not provide consistent sentential similarity ratings). Data sets will be collected incrementally, so that we can adjust their focus (for example, shifting focus phrase similarity ratings). WP2 models will be evaluated on the usable subset of empirical benchmarks, as well as by alternative evaluation means (e.g., by how well distributional functions approximate their output on held-out data).

b.2 Pilot study

The results of Baroni and Zamparelli (2010) show that it is possible to learn functions representing intensional adjectives such as "*different"*, that have virtually no content of their own independently of how they act on the nouns they modify. The step to determiners and other grammatical words is not huge. In a pilot study conducted for COMPOSES, we tested the distributional properties of noun phrases containing various determiners.

We built a simple distributional semantic space from the British National Corpus for 5K frequent nouns and of the noun phrases where these nouns were preceded by "*a", "the", "some", "any", "many",* and "*all"*. The space we built should capture broadly semantic rather than syntactic characteristics, since we recorded co-occurrence within a full sentence span (not only adjacent elements), we did not keep track of the relative position of collocates, and both the targets (the nouns and noun phrases) and context expressions were lemmatized, discarding inflectional information. We extracted the 10 nearest neighbours (by cosine distance) of each noun or noun phrase target, and we pooled the neighbours by determiner type, analyzing them, in turn, grouped by determiner type of the target. For example, we counted, among all the neighbours of all the "*some"* phrases, how many are simple nouns, how many are "*a"* phrases, etc. We stress that these are neighbours in distributional semantic space, i.e., they are expressions that tend to occur in contexts similar to those of the targets, not (necessarily) expressions that directly co-occur with the targets.

The very clear pattern emerging from the statistical analysis of the results is that, among the neighbours of phrases with a certain determiner, phrases with the same determiner are always statistically over-represented (e.g., "*the"* phrases tend to have more "*the"* phrases among their neighbour, etc.). Moreover, for the quantifying determiners ("*all", "any", "some", "many"*), the other quantifying determiners are also over-represented. Of course, different phrases sharing the same noun tend to have related neighbours, but these neighbours also reflect the influence of the determiner. For example, among the nearest neighbours of "*a boycott"* we find "*an opposition"* and "*an alliance"*, whereas among those of "*the boycott"* we find "*the ban"* and "*the rally"*. Thus, corpus-extracted DSRs of determiner-noun phrases are capturing similarities in the contexts not only of the nouns, but also of the determiners, making NPs with the same or similar determiner more similar to each other.

What makes the determiner-sharing DSRs similar? To answer this question, we compute the centroid DSR vector for each determiner (e.g., we compute the average DSR of "*some dog", "some cat"* and all other "*some"* noun phrases), and we extract the vector dimensions that are most significantly associated with each of these average DSRs. With the exception of "*the"*, that produces rather inconspicuous results, for all other centroids at least some of the characteristic dimensions are very reasonable: terms related to vagueness for "*a"* ("*someone", "something", "somewhere", "somebody"*); modal verbs for the quantifying determiners ("*all must", "any shall", "some may"*); degree expressions ("*too", "so"*) for "*many"*; *"include"* for "*all"*; "*person"* for "*any";* "*other*" for "*some*"; and so on.

To conclude, the pilot exploration of the DSRs of noun phrases with various determiners (i.e., the sort of DSRs that will constitute training targets in distributional function learning) indicates that, even when the source corpus is small and feature extraction trivial, there is enough information out there to capture some intuitively meaningful generalizations about the phrases of interest.

b.3 References

Abney, S. 2007. *Semisupervised Learning for Computational Linguistics*. CRC.

Aitchison, J. 1993. *Words in the Mind*. Blackwell.

Artstein, R. and M. Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34(4): 555-596.

Asher, N. To appear. *Lexical Meaning in Context*. CUP.

Bader, M. and J. Häussler. 2010. Toward a model of grammaticality judgments. *Journal of Linguistics* 46: 273-330.

Baroni, M. and A. Lenci. To appear. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics.*

Baroni, M. and R. Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. *Proceedings of EMNLP*: 1183-1193.

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

Baroni, M., 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.

Blackburn, P. and J. Bos. 2005. *Representation and Inference for Natural Language*. CSLI.

Bos, J. 2005. Towards wide-coverage semantic interpretation. *Proceedings of IWCS*: 42-53.

Chierchia, G. and S. McConnell-Ginet. 2000. *Meaning and Grammar.* MIT Press.

Clark, S. and S. Pulman. 2007. Combining symbolic and distributional models of meaning. *Proceedings of QI*: 52-55.

Clarke, D. 2007. *Context-theoretic Semantics for Natural Language.* PhD thesis, University of Sussex.

Coecke, B., M. Sadrzadeh and S. Clark. To appear. Mathematical foundations for a compositional distributed model of meaning. *Lambek Festschrift*.

Cohen, A. 1999. *Think Generic!* CSLI.

Cowie, J. and Y. Wilks. 2000. Information extraction. In R. Dale, H. Moisl and H. Somers (eds.), *Handbook of Natural Language Processing*. Dekker.

Croft W. and A. Cruse. 2004. *Cognitive Linguistics*. CUP.

Dagan, I., B. Dolan, B. Magnini and D. Roth. 2009. Recognizing textual entailment: Rationale, evaluation and approaches. *Natural Language Engineering* 15(4): 1-17.

Erk, K. 2009. Supporting inferences in semantic space. *Proceedings of IWCS*: 104-115.

Erk, K. and S. Padó. 2008. A structured vector space model for word meaning in context. *Proceedings of EMNLP*: 897-906.

Guevara, E. 2010. A regression model of adjective-noun compositionality in distributional semantics. *Proceedings of GEMS*: 33-37.

Hampton, J .1991. The combination of prototype concepts. In P. Schwanenflugel (ed.), *The Psychology of Word Meanings*. Erlbaum.

Hastie, T., R. Tibshirani and J. Friedman. 2009. *The Elements of Statistical Learning*. Springer.

Herdağdelen, A. and M. Baroni. To appear. The Concept Game: Better commonsense knowledge extraction by combining text mining and a game with a purpose. *AAAI Commonsense Knowledge Acquisition Symposium*.

Jackendoff, R. 2002. *Foundations of Language*. OUP.

Jurafsky, D. and J.Martin. 2008. *Speech and Language Processing*. Prentice Hall.

Koller, D. and N. Friedman. 2009. *Probabilistic Graphical Models*. MIT Press.

Landauer, T. and S. Dumais. 1997. A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2): 211-240.

Lund, K. and C. Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behaviour Research Methods* 28: 203-208.

MacCartney, B. and C. Manning. 2009. An extended model of natural logic. Proceedings of *IWCS*: 140-156.

Manning, C., P. Raghavan and H. Schütze. 2008 *Introduction to Information Retrieval*. CUP.

Margolis, E. and S. Laurence. 1999. *Concepts: Core Readings*. MIT Press.

McDonald, S. 2000. *Environmental Determinants of Lexical Processing Effort*. PhD thesis, University of Edinburgh.

Medin, D., R. Goldstone and D. Getner. 1993. Respects for similarity. *Psychological Review* 100: 254-278.

Mitchell, J. and M. Lapata. 2008. Vector-based models of semantic composition. *Proceedings of ACL*: 236-

244.

Mitchell, J. and M. Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*: online.

Montague, R. 1970. English as a formal language. In B. Visentini (ed.), *Linguaggi nella Società e nella Tecnica*. Comunità.

Munro, R., S. Bethard, V. Kuperman, V. Lai, R. Melnick, C. Potts, T. Schnoebelen and H. Tily. Crowdsourcing and language studies: The new generation of linguistic data. *NAACL-2010 Workshop on Creating Speech and Language Data With Amazon's Mechanical Turk*: 122-130.

Pustejovsky, J. 1995. *The Generative Lexicon*. MIT Press.

Rogers, T. and J. McClelland. 2004. *Semantic Cognition: A Parallel Distributed Processing Approach*. MIT Press.

Thater, S., H. Fürstenau and M. Pinkal. 2010. Contextualizing semantic representations using syntactically enriched vector models. *Proceedings of ACL*: 948-957.

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

Weskott, T. and G. Fanselow. 2008. Variance and informativity in different measures of linguistic acceptability. *Proceedings of WCCFL*: 431-439.

c. Resources (incl. project costs)

c.1 Personnel

Personnel involvement timeline

	M6	M12	M18	M24	M30	M36	M42	M48	M54	M60
Baroni (PI)	■	■	■	■	■	■	■	■	■	■
Zamparelli	■	■	■	■	■	■	■	■	■	■
Bernardi	■	■	■	■	■	■	■	■	■	■
Post-Doc 1		■	■	■	■	■	■	■	■	
Post-Doc 2		■	■	■	■	■	■	■	■	
PhD Student 1			■	■	■	■	■	■		
PhD Student 2					■	■	■	■	■	■

Senior staff:
Marco Baroni (PI, 67% on project for 5 years), **Roberto Zamparelli** and **Raffaella Bernardi** (both 20% on project for 5 years).

To be hired with project funds:
Post-Doc 1: advanced expertise in data-intensive computing and machine learning (especially graphical models); **post-Doc 2:** experimental expertise in behavioural methods (ideally including elicitation of language data); **PhD Student 1:** will focus on computational modeling; **PhD Student 2:** will focus on implications for theoretical linguistics.

External consultants:

- **Nicholas Asher** (Institut de Recherche en Informatique, Toulouse, France): leading expertise in formal semantics.
- **Stephen Clark** (University of Cambridge, UK): leading expertise in natural language processing, including machine learning, distributional models and compositionality.
- **Katrin Erk** (University of Texas at Austin, USA): leading expertise in computational semantics, in particular distributional models and connections with cognitive science.

c.2 The host institution

The CIMeC Language Interaction and Computation Laboratory (**CLIC**), directed by Prof. Massimo Poesio, was inaugurated in 2007 and is currently composed of 2 full professors, 4 tenured/tenure-track researchers (including the PI and team members), 6 post-docs, 9 doctoral students and about 10 other collaborators with various affiliations. Despite its recent foundation, CLIC is already one of the largest and most active Italian centers specializing in computational linguistics, with particular emphasis on the links of computational linguistics with theoretical linguistics and cognitive science. The laboratory is currently involved in the European project GALATEAS and in the large locally funded LiveMemories project, as well as in a number of smaller scale projects. CLIC is part of a wider network of centers focusing on human language, knowledge and related areas in the Trento region. CLIC has close research and teaching ties with the SIS lab of the UNITN Computer Science department, the HLT group at the Trento FBK institute and the Trento-based Laboratory for Applied Ontology of the Italian CNR. CLIC is a unit in the University of Trento Center for Mind/Brain Sciences (**CIMeC**), that started in 2007 to promote interdisciplinary research in the cognitive (neuro)sciences, and has been growing very rapidly – as of late 2010 it numbers 67 researchers and 35 doctoral students, including psychologists, neuroscientists, physicists, linguists and computer scientists. CLIC members actively interact with other CIMeC researchers and students, thus strengthening their interdisciplinary ties to cognitive science. CIMeC is part of the University of Trento (**UNITN**), that consistently ranks among the top Italian universities for teaching and research, and features the best national track record in attracting European funds (according to a 2009 survey), including a CIMeC ERC starting grant in 2010. Consequently, UNITN has a very experienced Research Support Office that helps PIs with all the phases of grant management.

c.3 Summary of costs

See Sec. c.1 for description of personnel. Equipment cost (server extension) detailed in WP1.1. Travel includes conference participation and consultant invitation, Other direct costs: Other is for final workshop organization. Subcontracting costs are for audits.

	Cost Category	**Year 1**	**Year 2**	**Year 3**	**Year 4**	**Year 5**	**Total (Y1-5)**
Direct Costs:	*Personnel:*						
	PI	34453.5	34453.5	34453.5	34453.5	34453.5	172267.6
	Senior Staff	14594.6	15571	17226.8	17226.8	17226.8	81845.88
	Post docs	55000	110000	110000	110000	55000	440000
	Students	0	18333.4	36666.7	36666.7	18333.4	110000.16
	Other	0	0	0	0	0	0
	Total Personnel:	104048	178358	198347	198347	125014	804113.64
	Other Direct Costs:						
	Equipment	16666.7	16666.7	16666.7	0	0	50000
	Consumables	0	2500	2500	0	0	5000
	Travel	0	11500	17000	5000	10000	43500
	Publications, etc	0	0	0	0	0	0
	Other	0	0	0	0	20000	20000
	Tot. Other Costs	16666.7	30666.7	36166.7	5000	30000	118500
	Tot. Dir. Costs:	120715	209025	234514	203347	155014	922613.64
Indirect Costs (overheads):	Max 20% of Direct Costs	24142.9	41804.9	46902.7	40669.4	31002.7	184522.73
Subcontracting Costs:	(No overheads)	0	3500	0	3500	3500	10500
Total Costs of project:	(by year and total)	144858	254329	281416	247516	189516	1117636.4
Requested Grant:	(by year and total)	144858	254329	281416	247516	189516	1117636.4

For the above cost table, please indicate the % of working time the PI dedicates to the project over the period of the grant:	**67%**

d. Ethical issues

ETHICS ISSUES TABLE

Areas Excluded From Funding Under FP7 *(Art. 6)*

(i) Research activity aiming at human cloning for reproductive purposes;

(ii) Research activity intended to modify the genetic heritage of human beings which could make such changes heritable (Research relating to cancer treatment of the gonads can be financed);

(iii) Research activities intended to create human embryos solely for the purpose of research or for the purpose of stem cell procurement, including by means of somatic cell nuclear transfer;

All FP7 funded research shall comply with the relevant national, EU and international ethics-related rules and professional codes of conduct. Where necessary, the beneficiary(ies) shall provide the responsible Commission services with a written confirmation that it has received (a) favourable opinion(s) of the relevant ethics committee(s) and, if applicable, the regulatory approval(s) of the competent national or local authority(ies) in the country in which the research is to be carried out, before beginning any Commission approved research requiring such opinions or approvals. The copy of the official approval from the relevant national or local ethics committees must also be provided to the responsible Commission services.

	Research on Human Embryo/ Foetus	YES	Page
	Does the proposed research involve human Embryos?		
	Does the proposed research involve human Foetal Tissues/ Cells?		
	Does the proposed research involve human Embryonic Stem Cells (hESCs)?		
	Does the proposed research on human Embryonic Stem Cells involve cells in culture?		
	Does the proposed research on Human Embryonic Stem Cells involve the derivation of cells from Embryos?		
	I CONFIRM THAT NONE OF THE ABOVE ISSUES APPLY TO MY PROPOSAL	X	

	Research on Humans	YES	Page
	Does the proposed research involve children?		
	Does the proposed research involve patients?		
	Does the proposed research involve persons not able to give consent?		
	Does the proposed research involve adult healthy volunteers?	X	
	Does the proposed research involve Human genetic material?		
	Does the proposed research involve Human biological samples?		
	Does the proposed research involve Human data collection?		
	I CONFIRM THAT NONE OF THE ABOVE ISSUES APPLY TO MY PROPOSAL		

	Privacy	YES	Page

	Does the proposed research involve processing of genetic information or personal data (e.g. health, sexual lifestyle, ethnicity, political opinion, religious or philosophical conviction)?		
	Does the proposed research involve tracking the location or observation of people?		
	I CONFIRM THAT NONE OF THE ABOVE ISSUES APPLY TO MY PROPOSAL	X	

	Research on Animals	YES	Page
	Does the proposed research involve research on animals?		
	Are those animals transgenic small laboratory animals?		
	Are those animals transgenic farm animals?		
	Are those animals non-human primates?		
	Are those animals cloned farm animals?		
	I CONFIRM THAT NONE OF THE ABOVE ISSUES APPLY TO MY PROPOSAL	X	

	Research Involving non-EU Countries (ICPC Countries)	YES	Page
	Is the proposed research (or parts of it) going to take place in one or more of the ICPC Countries?		
	Is any material used in the research (e.g. personal data, animal and/or human tissue samples, genetic material, live animals, etc) : a) Collected in any of the ICPC countries?		
	b) Exported to any other country (including ICPC and EU Member States)?		
	I CONFIRM THAT NONE OF THE ABOVE ISSUES APPLY TO MY PROPOSAL	X	

	Dual Use	YES	Page
	Research having direct military use		
	Research having the potential for terrorist abuse		
	I CONFIRM THAT NONE OF THE ABOVE ISSUES APPLY TO MY PROPOSAL	X	

If you have answered "YES" to any of the above questions you are required to complete and upload the "B2_Ethical Issues Annex" (template provided).

ANNEX SUBMITTED.