ERC Starting Grant 2016
Research proposal [Part B2]

Part B2: *The scientific proposal*

Section a. State-of-the-art and objectives

a.1 Objectives

Humans use language to communicate about the world, from immediate sensory information ("Caution, it's hot!") to very abstract knowledge acquired through the years or even through generations ("The Universe is expanding"). Modeling the process by which we use linguistic expressions to refer to the outside world is fundamental for understanding language, a defining trait of the human species. The goal of AMORE is to advance the state of the art in Computational Linguistics, Linguistics, and Artificial Intelligence by developing **a model of linguistic reference to entities implemented as a computational system** that can learn its own representations from data. I focus on concrete entities (physical objects humans perceive as a unit) because they constitute a well-delimited domain, representative of the larger reference problem (see Section a.4).

Linguistic reference crucially involves both **continuous** and **discrete** aspects of meaning. A noun phrase such as "the big tree" gives us some *descriptive content* that allows us to identify a particular entity through some of its properties (Frege 1892). However, this descriptive content is notoriously fuzzy (Fodor et al. 1980; Cruse 1986; Keefe 2000): The word "tree" applies to "many unlike individuals of diverse size and form" (Borges 1944), from near-bushes to sequoyas to even genealogical trees, with no definite criteria nor clear boundaries between what counts as a tree and what doesn't. Fuzziness persists when composing words into phrases and sentences: For instance, "red car" applies to objects with a different color than "red cheek", and the semantic contribution of "red" changes in a continuous fashion depending on the modified noun (Boleda et al. 2013), e.g. going towards pink or orange. The fuzzy, continuous nature of meaning is actually a very useful trait of language, and the conceptual system it relies on, because it allows us to handle an infinitely varied, ever-changing reality reusing knowledge about previously encountered situations (Murphy 2002; van Deemter 2010).

When phrases with fuzzy descriptive content are used in a specific context, however, they are used to refer to specific entities in the real world. Humans treat the *referents* picked out by referential expressions as essentially discrete, and language offers us tools to deal with that, too. For instance, a speaker uttering example (1) uses the noun phrase "a box" to introduce a specific, discrete entity in the current discourse, and the anaphoric pronoun "it" to refer back to precisely that entity and add more information about it (Kamp and Reyle 1993).

(1) The man lifted a box. It was heavy.

Thus, linguistic referents are both discrete (they are clearly *delimited* with respect to each other and *individuated* through linguistic mechanisms such as the use of noun phrases and pronouns) and continuous (they are linked to rich descriptive content). Some of the most successful previous work in theoretical and computational semantics, however, is markedly biased towards one aspect, at the expense of the other. *Formal semantics* (Montague 1970 and subsequent work) employs logic and other symbolic mathematical tools to provide discrete semantic representations of linguistic expressions. This approach has advanced our understanding of the linguistic mechanisms that individuate referents. For instance, it can model the fact that when I say "the big tree" I pick out a unique object that is a big tree, that indefinite noun phrases ("a box") are used to introduce new referents in the discourse whereas definite ones ("the man") point to already accessible referents, and that pronouns are pointers to referents (Kamp 1981, Kamp and Reyle 1993). However, formal semantics says little about the characteristic properties of a big tree or a box, and consequently about how we are able to pick out the right entity in the first place. *Distributional semantics* (see Turney and Pantel 2010 for an overview) models meaning in terms of context of use, and it provides numerical, continuous distributed representations for linguistic expressions. This approach focuses on descriptive content, successfully modeling graded semantic phenomena such as word and phrase similarity ("box" – "package", "important route" – "major road"; Landauer and Dumais 1997; Baroni and Zamparelli 2010) and meaning modulation (the difference in red in "red car" vs. "red cheek"; Boleda et al. 2013). However, current distributional models do not have a notion of an individuated referent to attach the descriptive content to: Their semantic representations do not have the right structure to distinguish between different referents. As a result, they might for instance tell us that "box" and "package", being conceptually

similar, can in principle denote the same referent, but they are useless at telling us if, in a specific discourse, the two expressions are indeed pointing to the same referent or not.

The shortcomings of each approach severely limit their applicability in Computational Linguistics tasks that require natural language interpretation. For instance, Bos and Markert (2005) applied a formal semantic system to Recognizing Textual Entailment, which seeks to model natural language inference (e.g., "Crude oil prices soared to record levels" entails "Crude oil prices rose", but "The white man spoke" does not entail "The black man spoke"). The system was reasonably precise: When it predicted entailment, it was right 77% of the time. However, it was able to predict only 6% of the entailments; because of its poor treatment of descriptive content, it had very low coverage. This system was for instance unable to relate "[a] sport utility vehicle drifted onto the shoulder of a highway and struck a parked truck" to "car accident". Systems that solely use distributional semantics have a better coverage but a lower precision (Beltagy et al. 2013): Entailment decisions require predicting whether an expression applies to the same referent or not, and these systems make trivial mistakes, such as deciding that "white man" and "black man", being conceptually similar, should co-refer.

I believe that the lack of tools to adequately handle descriptive content is an insurmountable roadblock for symbolic approaches such as formal semantics. My strategy therefore is to take inspiration from some important insights in the formal tradition, but focus on pushing the limits of distributional semantics, a fast-moving and exciting area to which I am strongly contributing. AMORE will **endow distributional semantics with referential capabilities**. Thanks to my original adaptation to language of very recent developments in Machine Learning, the AMORE model will be able to automatically induce and operate with **individuated referents** which, however, have **a continuous distributed internal representation**, accounting for the conceptual richness of the process of referring. The use of these advances further allows the model to integrate two major sources of information about referred entities (Kamp 2015): **Perceptual** information from the environment, and **previous knowledge** gained through language.

The goals of AMORE are to:

- Develop a model of reference to entities that links descriptive content and individuated referents.
- Implement it as a computational (specifically, neural network) system that is able to learn representations from data.
- Test it extensively in experiments involving reference to entities in discourse and in the perceptual (visual) environment.
- Explore the consequences of moving to distributed entity representations for Computational Linguistics, Linguistics more generally, and Artificial Intelligence.

a.2 The project in the context of the state of the art

Here we explain how the project stands with respect to previous theoretical and computational work that is of direct relevance to it. We will not review some important approaches to semantics, such as cognitive linguistics (Croft and Cruse 2004) or the psychological and philosophical literature on concepts (Margolis and Laurence 1999), although some of their concerns and empirical results will be present in the project.

Formal semantics (Montague 1974 and subsequent work) is rooted in analytic philosophy. It uses symbolic mathematical tools, prominently logic, to represent natural language meaning, with a special emphasis on the effect of semantic composition. Formal semantics puts reference center stage: It treats meaning as a mapping from linguistic expressions to the world. Typically, the referents of expressions are represented as variables in a logical representation. Having explicit variables for referents is very useful: Among other advantages, it makes it easy to track referents as a discourse unfolds, and it also allows for systematic links between the shape of linguistic expressions and their referential import, as explored in Discourse Representation Theory (DRT), a prominent formal semantic framework that inspires our distributional model (Kamp 1981; Kamp and Reyle 1993). For instance, a simplified formalization of the two-sentence discourse example (1) above in DRT would be as follows:

(2)

x y
man(x)
box(y)
lifted(x,y)

(The man lifted a box.)

(3)

x y
man(x)
box(y)
lifted(x,y)
heavy(y)

(It was heavy.)

In (2), we find the semantic representation of the ongoing discourse after processing the first sentence: There are two discourse referents, represented by variables *x* (for the entity denoted by "the man") and *y* (introduced by the indefinite noun phrase "a box"). In (3), we augment the representation with the

information in the second sentence. The sentence concerns an entity that, as signaled by the anaphoric pronoun "it", is already introduced; thus, after resolving the anaphora, we obtain a representation that shows that the entity that is a box, lifted by the man, and heavy is one and the same. DRT specifies a systematic mapping from natural language sentences to this kind of semantic representation, and it has been implemented in a computational tool that produces DRT representations for free English text (Bos 2008).

Words with descriptive content, such as nouns, adjectives and verbs, are represented as logical predicates: for instance, "box" is translated as a predicate that applies to a single argument (in the example, the discourse referent represented by *y*). These predicates have no internal structure nor link to their descriptive content (beyond the mapping to their referents; but then, without an account of the descriptive content it is not clear how speakers can connect the predicates to their referents in the first place; Baroni et al. 2014a). As a result, for instance the words "box" and "package" correspond to distinct predicates, as distinct as, say, "box" and "smile". There have been attempts to address this problem within formal semantics, such as the use of meaning postulates and decompositional mechanisms (Montague 1960; Katz and Fodor 1963; Pustejovsky 1995; Asher 2011), but these have been only partially successful (Fodor et al. 1980; Murphy 2002; Boleda and Erk 2015). In AMORE, I develop a model that, like formal semantics approaches, can individuate and identify referents, but also associate them with rich descriptive content.

Distributional semantics (see Turney and Pantel 2010 for an overview) is based on the hypothesis that the meaning of a linguistic expression can be induced from the contexts in which it is used (Firth 1957; Harris 1968; with possible roots in Wittgenstein 1953), because related expressions, such as "box" and "package", are used in similar contexts ("open the _", "a light _"). This provides an operational learning procedure for semantic representations that has been profitably used in computational semantics.

Distributional representations for meaning are *vectors* (essentially lists of numbers; they can also be more complex algebraic objects such as matrices and tensors). Vector values are abstractions on the contexts of use, extracted from large amounts of natural language data such as web text. The semantic information is *distributed* across all the dimensions of the vector, and it is expressed in the form of *continuous* values, which allows for rich and nuanced information to be encoded. The collection of linguistic elements, for instance words in a lexicon, forms a vector space or *semantic space*, in which semantic relations can be modeled as geometric relations. Thus, in semantic space, "box" is near to "package", and far from unrelated words such as "smile". In recent years, distributional models have been extended to handle the semantic composition of words into phrases and sentences (Mitchell and Lapata 2010, Socher et al. 2013, Baroni et al. 2014a). Although these models still do not account for the full range of composition phenomena that have been examined in formal semantics, they do encode relevant semantic information, as shown by their success in demanding semantic tasks such as predicting sentence similarity (Marelli et al. 2014).

Current distributional models do not have a notion of referent, of a language-external entity that semantic representations are anchored to. This has adverse effects, such as the difficulty in distinguishing between general semantic relatedness and entailment discussed above. Two paths to overcome these difficulties have been recently explored, in both of which I have been involved: (1) Keep the logic framework and enrich it with distributional knowledge (Beltagy et al. 2013; Lewis and Steedman 2013); (2) Keep the distributional framework and anchor it in the external world (Gupta et al. 2015; Herbelot and Vecchi 2015). In AMORE I pursue approach (2).

Another research direction in which distributional semantics is advancing, and in which I have also participated, is to connect perceptual information to linguistic representations, e.g. by using visual cues extracted from images to improve the representation of color terms (Bruni et al. 2012). Distributional semantics provides an integrated representation of information from different modalities: linguistic (extracted from text), visual (from images), acoustic (from audio files), etc. This strand of research, like AMORE, addresses the issue of grounding symbolic language in external reality (Searle 1980; Harnad 1990), though not specifically reference. AMORE makes use of advances in this area to model **linguistic reference to objects apprehended through perception**, matching noun phrases containing visual attributes ("the white cat") with entities depicted in an image (WP 4.1). It also enables it to **integrate information coming from perceptual and linguistic cues**: For instance, we expect the model to match "the sociable white cat" to an image of a white cat for which the only linguistic information we have given is "is sociable" (WP 4.2). In the project we focus on visual information because (1) visual information is crucial for concrete entities (the domain of the project), (2) Computer Vision is mature enough that incorporating it into our computational model of reference is feasible (see below).

Neural networks, especially in the variant called *deep learning* (LeCun et al. 2015), are a Machine Learning algorithm developed in the last century (Rosenblatt 1958 and subsequent work) that is receiving renewed interest due to significant breakthroughs in many Artificial Intelligence areas, such as Computer Vision (Krizhevsky et al. 2012), Speech Processing (Hinton et al. 2012), and Computational Linguistics (Mikolov et

al. 2013). Machine Learning algorithms are used to learn computational models from data to perform tasks that can be applied to new, unseen cases (Witten and Frank 2005). Learning consists in setting the values of some pre-specified parameters in the learning algorithm. Traditional supervised Machine Learning methods needed careful, expert-driven extraction of features, that is, informative elements, from the data,. Neural networks, in contrast, belong to the *representation learning* paradigm that seeks to automatically induce the right features from the data. Deep neural networks can learn very complex functions from input (say, the pixels of an image) to output (e.g. the name of the object it depicts) by composing successive transformations of the representations (see Nielsen 2015 and LeCun et al. 2015 for algorithmic details). These transformations are continuous in nature, which allows for them to be learned by using advanced optimization techniques; however, they can approximate discrete operations (see discussion of the operations on the entity representations in Section a.3).

Distributional semantics has recently started using neural networks as a powerful tool to learn its semantic representations, consistently improving results across a wide range of semantic benchmarks (Baroni et al. 2014b). Recall that in distributional semantics the representations are based on abstraction operations on linguistic contexts. There is a wide range of possible abstractions; the main advantage of neural networks is that the specific operation to perform on the contexts is based on a prediction task. This way, vectors are initialized randomly, and they are iteratively refined as the model goes through the data and improves its predictions. Linguistic prediction tasks that are general enough lead to general-purpose semantic representations; for instance, Mikolov et al. (2013) used language modeling, the task of predicting words in a sentence (e.g. "She lifted her cup and took a _"). The result is one of the highest quality distributional lexicons available today. Neural networks also facilitate the use of *transfer learning*, or transferring knowledge from one learning task to another, by taking advantage of the features they automatically induce. For instance, Weiss et al. (2015) use Mikolov et al. (2013)'s word vectors to initialize the word representations for a syntactic task. Given that syntactic datasets are small, this gives the model a good start (in other approaches, all the knowledge needs to be learned from scratch for each new task). I will exploit similar initialization strategies in AMORE. More importantly, neural networks will play a key role in making it possible to produce distributed representations of the entities that we refer to via language, the endeavour I will pursue in AMORE.

Specifically, AMORE adopts Recurrent Neural Networks (RNNs; Elman 1990), which are especially suited for linguistic tasks because they adequately handle sequential input. In RNNs, when processing a given word, the network uses the representation it has built of the previous discourse. For instance, given the sentences in (1) above, the network can use the representation built for "The" when predicting "man", "The man" when predicting "lifted", etc. In the process of learning to predict the following word in the training corpus (or other distributional tasks), a RNN will induce a vector-based representation of a word (known as an embedding in the neural network literature) that is substantially analogous to the word vectors constructed with traditional distributional semantic methods. Thus, RNNs can be naturally seen as extensions of word-based distributional semantic approaches that also account for semantic and syntactic properties of broader constituents.

Note in particular that the discourse representation that a RNN sequentially builds is the result of a simple form of semantic composition (Li et al. 2015). Linguistic structures, however, are not only sequential but also hierarchical; accordingly, they are often represented using syntactic trees (Chomsky 1957). RNNs, especially in their LSTM variant (Hochreiter and Schmidhuber 1997), can partially model hierarchical structures and thus syntax, through a gating mechanism that allows different elements in the previous discourse to have different effects on the representations at different times, resulting in better handling of long-distance dependencies (in the remainder of this proposal, when we refer to RNNs, we always imply the LSTM variant). Explicitly hierarchical models, *Recursive* Neural Networks, are also being exploited for language (Socher et al. 2013, Le and Zuidema 2014), but they do not at present show consistent improvements over RNNs (Li et al. 2015) and the latter are easier to integrate with the dynamic memory mechanisms that we will discuss next.

A RNN induces a single continuous representation of the semantic information active at each time step in the unfolding discourse, with nothing akin to the distinct variables used in formal semantics to represent entities. The crucial technical innovation of AMORE will be to extend RNNs with a dynamic memory that the network can learn to manipulate, storing distinct representations in it, and retrieving them when needed. This allows us to emulate **discrete operations on entity representations in a continuous setup**. The specific mechanisms build on recent work that uses continuous approximations to discrete operations such as storage and retrieval, in order to integrate them in architectures that, being fully differentiable, can learn from data through effective methods (e.g., Joulin and Mikolov 2015, Sukhbaatar et al. 2015, Bahdanau et al. 2015).

Computational Linguistics and Artificial Intelligence tasks and resources. The Computational Linguistics and Artificial Intelligence communities have tackled tasks that are related to the goals of AMORE, and developed resources that will be useful for the project.

- *Coreference* or *anaphora resolution* (Poesio et al. in press) involves identifying linguistic expressions that have the same referent. It has often been operationalized as follows: Given a text, (1) identify all the expressions with referential import ("mentions"); and (2) determine which of them have the same referent. The full coreference task involves a number of issues that are not central for the goals of this project, such as being able to explicitly recognize all the mentions of a given referent. However, we will capitalize on the theoretical and empirical evidence gathered in this area of research, and use it to evaluate the model in an extrinsic task. In particular, we will test our model in a subset of OntoNotes, a corpus annotated for coreference (Hovy et al. 2006).
- A number of related tasks address the ability of computational systems to understand written text (*Machine Comprehension of Text*, Richardson et al. 2013). Neural network systems are starting to be tested on this kind of task, and large data sets are being developed (Hermann et al. 2015; Hill et al. 2015). *Question Answering* (Voorhees 1999) involves asking systems to answer questions in natural language, which can be generic ("Where is the Taj Mahal?") or about a specific document, when applied to text understanding. AMORE aims at connecting linguistic expressions with referents, and one of our experiments (WP3.1) can be framed as a language-understanding question answering task regarding entities in text. Because it is a novel task, no existing resource is adequate to address it. By partially reusing existing methodology (Richardson et al. 2013), AMORE will create a comprehensive dataset for this task that will be very useful for the community.
- *Object recognition* is a task in Computer Vision that asks systems to identify and label objects in images ("cat", "plant", etc.; Krizhevsky et al. 2012). Recent work extends the task to recognizing attributes (such as *white* for cat; Russakovsky and Li 2012), and also addresses visual Question Answering (Antol et al. 2015). The state of the art in object recognition is mature enough that we can integrate an image processing component to model linguistic reference to objects depicted in images (WP4). We will also exploit existing Computer Vision data sets in our experiments: (1) a data set with images of single objects annotated with visual attributes (Russakowski and Li 2012; WP4.1); (2) the ReferItGame dataset, with human-generated referring expressions for objects in images (Kazemzadeh et al. 2014; WP4.2); (3) and possibly existing caption datasets (Young et al. 2014; Chen et al. 2015) for transfer learning. We will also reuse part of the methodology used for the ReferItGame dataset for the WP3.1 dataset.

a.3 The model

The AMORE model is essentially a distributional version of Kamp (2015), based on Discourse Representation Theory, which is one of the most comprehensive theoretical models of the interpretation of noun phrases to date. Kamp's model has four main components, accounting for: (1) generic information about the world, such as the fact that books have covers (K_{gen}); (2) information about the immediate extralinguistic environment (K_{env}); (3) the current linguistic discourse (K_{dis}); (4) the entities we talk about (K_{enc}, for "encyclopedic"). Interestingly, Kamp explicitly declares K_{gen} "off-limits" for his model (Kamp 2015, p. 54), consistent with the fact that it concerns descriptive content.

Figure 1 depicts the model to be developed in the project. Like Kamp's, it is a model of language interpretation (as opposed to generation), and it simulates the situation in which a competent hearer is processing an ongoing discourse.[1] It has four main components, labeled with letters in the figure. The first component (A) processes the *linguistic* input, and it includes a distributional word lexicon (induced during model training) and a recurrent composition layer;[2] component (B) represents the perceptual *environment*; and component (C) handles the *entities* in the discourse and the environment. The fourth component (D) integrates visual and linguistic information and transfers information to and from the entity library. The K_{gen} component of Kamp's model is operationalized in AMORE as the corpus-induced distributional lexicon, because distributional representations have been shown to account for generic information (Baroni and Lenci 2010, a.o.). K_{env} and K_{enc} correspond to our environment component and entity library, respectively. Both in Kamp's and in our model, the entity library contains representations for previously known entities (if I say to my mother "Aunt Lina came today", she can access her long-term entity representation for her sister) as well as those newly introduced in a discourse ("I just saw *a bird*"). This project focuses mainly on newly

[1] The model only encodes the contents of what is being said, without keeping track of what was said when, and so does not does not cover aspects that are sensitive to information structure (Vallduví and Engdahl 1996). This is left to future work.

[2] Layers are the successive representations used by the network: Networks go from input to ouput through a series of hidden layers that correspond to intermediate, progressively more abstract representations.

introduced entities (Work Packages 3.1, 4.1 and 4.2), but one of the experiments is targeted at previously known entities (WP 3.2). The information in K_{dis} is distributed in AMORE between the composition module, the integration module, and the entity library.

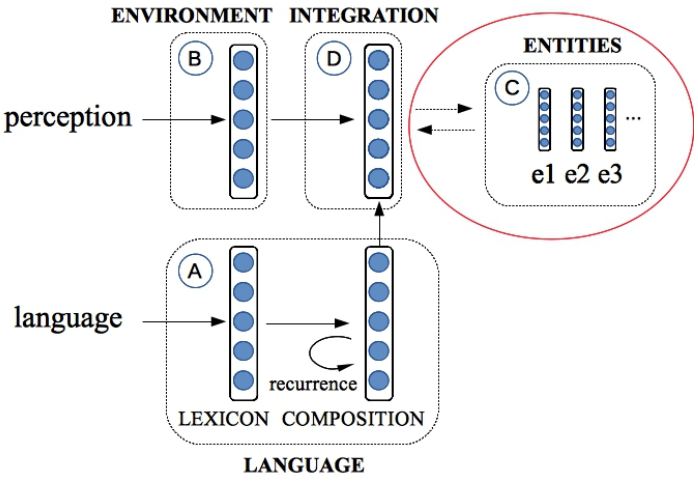

Figure 1. The AMORE model.

The model is implemented as a Recurrent Neural Network (with standard LSTM gating mechanisms to keep track of longer dependencies), extended by a dynamic memory storing the entity library and an additional hidden layer for the integration component. The environment component is a neural network itself (in practice, we will use the pre-trained convolutional neural network of Krizhevsky et al. 2012). Our network processes one piece of input at a time (e.g., each word in a text, each image in a sequence of images; the exact timing in which images and words are presented will depend on the task). The operations mapping the visual and linguistic input to the integration layer are standard matrix multiplications followed by nonlinearities (Nielsen 2015). The main novelty of the model is the use of the dynamic memory for the entity library and the mechanisms to create, update, and retrieve distributional entity representations. Without the entity library, the integration layer would have to represent multiple entities all squashed into the same distributed representations, whereas we need discrete pointers to make sense of phenomena such as anaphora. The mechanisms we propose are inspired by very recent progress in neural networks, enabling them to manipulate external memories and retrieve/focus on specific memory elements while maintaining a fully differentiable architecture, i.e., one that can learn from data using standard gradient-based techniques (e.g., Joulin and Mikolov 2015, Sukhbaatar et al. 2015, Bahdanau et al. 2015). Figure 2 zooms in on this part of the model, marked with a red ellipse in Figure 1. In Figure 2, continuous lines again denote matrix multiplications followed by non-linearities, whereas dashed lines indicate the input/output flow from the modules described in detail below (what follows is just one possible implementation of the relevant mechanisms, and AMORE will explore several variants).

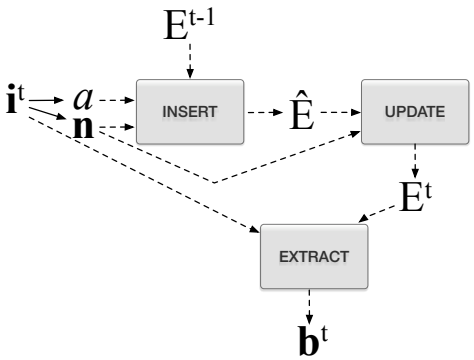

Figure 2. Operations on the entity library.

The entity library is continuously modified as the input is processed. We start from vector $\mathbf{i}^t$, produced by the integration component by collating information from the current linguistic input, the visual input (if any), and the information in the entity library right before the current input was read. The modification to the entity library is carried out in two steps: First, INSERT adds an element to the library, if necessary; second, UPDATE updates the whole entity library with the new information. The INSERT module, inspired by Joulin and Mikolov's (2015) Stack RNN, uses two pieces of information: A vector **n**, that should encode the new relevant information that the input added to the discourse, and a "controller" scalar value *a* ranging between 0 and 1 that will determine the effect of the new information on the entity library.[3] INSERT takes as input *a*, **n** and the entity library in its current state (E^{t-1})[4] and decides whether to "softly" insert **n** as a new entity representation at the top of the library, shifting all other vectors by one position, according to Equations (1) and (2) below. As *a* approaches 1, the operation converges to discrete insertion of a new element, whereas *a* approaching 0 means that the entity library will not be changed (note that, as discussed above, this and the next operations approximate discrete actions by continuous means to make learning possible).

INSERT builds a new, temporary version of the entity library (Ê). UPDATE uses Ê and (again) vector **n** to update the library, arriving at the final representation of the library at the current time step, E^t. The UPDATE module decides how much to update each entity vector with the new information in **n**. A

[3] Scalar *a* is sigmoid-transformed so that it ranges between 0 and 1.
[4] E^{t-1} is a matrix with the entity vectors as columns, denoted by subscripts in the equations below.

probability vector **p** the size of the library is first obtained through Equation (3).[5] This will result in larger probability values for entity vectors that are similar to **n** (ideally, the probability mass should be largely concentrated on a single vector, referring to the entity we are talking about). The probabilities are then used to weight how much of the new information should be added to each entity vector in the updated library, according to Equation (4).[6] Finally, an EXTRACT operation produces a vector $\mathbf{b}^t$ (for "background") summarizing the information in the entity library, which will be used in the next input reading step as part of the information fed to the integration layer. EXTRACT is inspired by differentiable retrieval mechanisms such as those in Sukhbaatar et al. (2015), and produces $\mathbf{b}^t$ by summing across information in the current entity vectors while giving more weight to those that seem currently more relevant (relevance is operationalized as similarity to $\mathbf{i}^t$); see equations (5) and (6). This allows the library to feed back to the integration component. The same mechanisms operate on visual input as well (recall that the integration component encodes both linguistic and visual information).

(1) $\hat{\mathbf{e}}_1 = a\ \mathbf{n} + (1\text{-}a)\ \mathbf{e}_1^{t-1}$

(2) $\hat{\mathbf{e}}_i = a\ \mathbf{e}_{i-1}^{t-1} + (1\text{-}a)\ \mathbf{e}_i^{t-1}$ (for i>1)

(3) $\mathbf{p}=(softmax(\mathbf{n}^T\hat{E}))^T$

(4) $\mathbf{e}_i^t = tanh(\hat{\mathbf{e}}_i + p_i\ \mathbf{n})$

(5) $\mathbf{q}=(softmax(\mathbf{i}^{tT}E^t))^T$

(6) $\mathbf{b}^t=E^t\mathbf{q}$

For some intuition on how the model might work, consider input "a bird". At time 1, the network processes the indefinite article "a" and the entity library is empty. INSERT should create a new entity, because the network can be expected to learn to associate indefinite determiners, which introduce new discourse referents, with a high value for *a*. UPDATE should not touch the rest of the entity library, because determiners do not introduce any descriptive content. At time 2, when processing "bird", INSERT should not create a new entity, but UPDATE should modify the previously created one, because nouns contribute descriptive content. If the phrase occurs within a discourse, UPDATE can also modify any other entity representations that may be related to the "bird" entity, either explicitly in the input or through implicit connections (e.g., encoded in the distributional lexicon).

The architecture just specified needs appropriate data to learn from. Note in particular that the entity library will not be manually encoded, but it will have to be constructed by the model motivated by the tasks it is faced with. The model is trained on tasks that encourage it to create and access entity representations (see WPs 3 and 4 below). For instance, in WP3.1 the task is to predict the name of the character in a story that corresponds to a given noun phrase: Given a story with four characters, Ann, Bob, Tom, and Mary, that starts "Bob is a Law student, but Ann studies Medicine", the system should output "Ann" in response to "The Medicine student". The specific implementation of the output-producing mechanism depends on the task (see Section b.2).

The AMORE model is **able to keep entities distinct** (they correspond to different vectors in the library) **while at the same time providing rich internal distributed representations for them**, since the descriptive content and perceptual features associated with the entity are stored in its distributed representation. We can thus capitalize on the power of distributional semantics to handle different linguistic expressions that are semantically related ("box" - "package") and integrate linguistic and perceptual information (linking noun phrases like "the white cat" to entities depicted in images); at the same time, we also emulate symbolic variables by letting the network write and read from the entity library. This is a highly novel approach to entity representation and the first linguistically motivated use of neural networks with dynamic memory structures.

a.4 High-risk high-gain nature of the project and feasibility

AMORE is a radically new approach to reference with high-gain potential, because it provides a unified, scalable way to deal with the continuous and discrete aspects of reference, keeping track of distinct entities while providing rich conceptual representations for them. It is also high-risk, specifically with respect to the following points: (1) The model is completely data-induced, with little control (beyond specifying its architecture) over what each component does. This is what makes it powerful, since it learns the representations it needs to solve the task at hand. However, this also makes it risky. In particular, there is no guarantee that the entity library will effectively store entity representations, as opposed to some other type of

[5] The *softmax* function returns a probability distribution. The T superscript here and below denotes transposition operations that are needed for the dimensionalities of different vectors and matrices to match.

[6] The *tanh* function insures that entity vectors are kept within the same range.

information. (2) While learned continuous distributed representations have been shown to be very effective representations of both words and more complex linguistic expressions, there is no significant previous experience in encoding discourse entities as vectors (Kalchbrenner and Blunsom 2013 and Ji and Eisenstein 2015 apply neural networks to discourse tasks, but they do not use distributional entity representations like ours).

Given its inherent high-risk nature, I have carefully designed the project so as to maximize feasibility: (1) Given that the reference problem is huge, I have selected a well-defined domain with a solid theoretical framework to rely on: Reference to concrete entities (excluding, for instance, abstract entities and events), focusing on single-entity denoting noun phrases (excluding, for instance, plurals and group nouns); (2) I have designed tasks that will encourage the model to build entity representations and to match them with the representations obtained from the other components; (3) In the model, I am using components that have previously been shown to work for tasks related to AMORE, in my research as well as that of others; (4) The experiments include simpler and more complex challenges, and will be informative even in case of partial success; (5) I will do extensive analysis at each point of the project to address potential shortcomings, reverting to more controlled architectures or tasks if necessary. For instance, the INSERT operation could be enabled only when encountering noun phrases, which can be identified with standard Computational Linguistics tools; or more effort could be devoted to the controlled domain task in WP4.1 if the task proves too hard to move to the open domain (see Section b.2).

Finally, even if the AMORE model itself fails, because the project poses challenges to the community that need solving and operationalizes them in a way that is at the limits of the state of the art, its results (data sets, modeling results) will be very valuable for the community. For instance, if the entity library stores other information, it could be revealing to analyze what it stores; in my previous experience, by using new tools in semantics I have encountered phenomena that traditional linguistic methods had not uncovered (to give just one example, a computational study, Boleda et al. 2013, led to the discovery of differentiated referential and conceptual effects in semantic composition, McNally and Boleda 2015). Similarly, if the distributed approach to entity representation does not work, it will be very useful for the AI community to know why and how it compares to symbolic alternatives.

My previous research experience also supports the feasibility of the project. I am in a privileged position to propose a project building on formal and distributional semantics, since I have contributed top research in both distributional semantics (Boleda et al. 2004, Mayol et al. 2005, Boleda et al. 2007, Boleda et al. 2012b,c,d, Bruni et al. 2012, Boleda et al. 2013, Roller et al. 2014, Boleda and Erk 2015, Gupta et al. 2015) and formal semantics (McNally and Boleda 2004, Boleda et al. 2012a, Arsenijević et al. 2014, McNally and Boleda 2015). I have carried out extensive research on specific topics that are relevant for AMORE, such as adjectival and nominal semantics and, recently (as the PI of a Marie Curie project), on comparing symbolic and distributed semantic representations, integrating visual and linguistic information, and extracting referential information from distributional vectors. I have also shown my leadership capabilities, for instance encouraging the cross-fertilization between formal and distributional semantics as a guest editor for a special issue on the topic in the top journal in Computational Linguistics and leading the development of computational linguistic resources (see CV).

a.5 Expected impact

AMORE is a highly interdisciplinary project that will bring Linguistics, Computational Linguistics, and Artificial Intelligence forward. The proposed model handles reference, as does formal semantics, but in a framework that can adequately deal with descriptive content, integrates perceptual and linguistic information, and can learn from data and so has a broad coverage. These advances will contribute to our scientific understanding of language, a defining trait of the human species.

The project also substantially contributes to the far-reaching, decades-long debate in Artificial Intelligence and Cognitive Science over *symbolic* vs. *distributed* approaches to cognition (Smolensky 1987, Fodor and Pylyshyn 1988, Churchland 1998, Fodor and Lepore 1999, LeCun et al. 2015, among many others) with a proposal that synthesizes strong aspects of both approaches. The specific proposal will impact Artificial Intelligence as the first linguistically interesting application of Artificial Intelligence algorithms that have until now been used either for toy tasks (Joulin and Mikolov 2015) or without a clear linguistic motivation (Weston et al. 2014). From a more applied perspective, although the project itself focuses on fundamental research, it paves the way for technologies that will enable machines to talk to us in situated applications.

AMORE will also have a tremendous impact on my own research career. I will be joining the Universitat Pompeu Fabra in the fall of 2016 as an assistant professor in a tenure-track position. The project will enable me to create my own research group and will represent a key step on the path to a permanent position. In turn, I will significantly contribute to the already strong profile of the University in Linguistics, Artificial

Intelligence, and Cognitive Science, and to the European science space more generally: In addition to the project's results, AMORE will open up future research and funding opportunities across the solid network I have established up to now, encompassing Italy, Germany, and the US, with stable collaborations on topics related to the present project (see CV).

Section b. Methodology

b.1 Methodological framework

AMORE will follow the methodology of empirical Computational Linguistics, which uses Machine Learning methods to learn models from data that can be applied to new, unseen cases (Manning and Schütze 1999). We operationalize each experiment (see Work Packages 3 and 4 below) as a Machine Learning task. For instance, in WP3.1, given a story with four characters, Ann, Bob, Tom, and Dean, that starts "Bob is a Law student, but Ann studies Medicine (…)", the system has to identify the character referred to by the phrase "the Medicine student" and give its name as output, "Ann". Our model is *supervised*: It learns from examples of correct responses for story / noun phrase pairs. Learning consists in setting the values of some pre-specified parameters. These parameters are induced from a training set, further adjusted in a development set, and evaluated on a separate test set to ensure that the model is indeed able to generalize to unseen data. In our example, the system is trained on a set of stories, phrases, and character names, and tested on a separate set of stories and phrases. We will use standard evaluation metrics such as accuracy (percentage of correct responses over the total) to assess the quality of our model.

We implement the architecture of our model as a *neural network* (see Section a.2). The parameters to be learned from data are essentially the ones that determine the successive transformations of the representations from input to output through each of the layers in the network (see Nielsen 2015 and LeCun et al. 2015 for details), the operations manipulating the entity library, and the entity representations (see Section a.3). Other parameters, such as the dimensionality of the distributed representations, will be tuned on the development set. Parameter values will be initialized either randomly or using transfer learning, providing high-quality initial representations for linguistic and visual input, which will be adapted to the reference phenomenon by training on the specific tasks of the project (more details in the workplan description, Section b.2). As standard in Machine Learning, as part of the experiements we will also implement baseline architectures to asses the contributions of our system and compare to alternative approaches from the literature.

For the construction of the dataset in WP3.1, we will use *crowdsourcing*, a method that enables researchers to elicit data from paid volunteers in the online community, and that is widely used for the construction of Computational Linguistic datasets (Marelli et al. 2014, Bowman et al. 2015, a. o.; for more information and ethical aspects, see the Ethical Self-assessment Annex).

For data and error analysis, we will use statistical methods as well as visualization techniques, including model ablation (leaving one of the components of the model out, to see its impact on the overall performance), evaluations on focused subsets (e.g. entities belonging to different categories, like people or locations), and visualizations of distributed representations such as heat maps or 2D and 3D projections (e.g., van der Maaten and Hinton 2008). A further crucial component of data and error analysis, both for data set construction and modeling, will be the manual examination of samples. We will iteratively refine data set construction based on manual inspection of datapoints produced with each improved design.

b.2 Workplan

Timeline

	M6	M12	M18	M24	M30	M36	M42	M48	M54	M60
WP1: Release of project output and infrastructure										
WP1.1: Computational infrastructure										
WP1.2: Release of code and data										
WP2: Model development										
WP2.1: Model definition and implementation										
WP2.2: Task adaptation										
WP3: Linguistic knowledge-based reference										
WP3.1: Reference to newly-introduced entities										
WP3.2: Long-term modeling of entities										
WP4: Perception-based and integrated reference										
WP4.1: Controlled domain										
WP4.2: Open domain										
WP5: Management and dissemination										
WP5.1: Project management										
WP5.2: Dissemination										

Work Package 1: Data and infrastructure [Months 1-60]

Goals: Set up computational infrastructure; release research output for replication and further research.

WP1 Activity 1: Computational infrastructure. (Lead: PI; support from university ICT staff) [Months 1-60]
In order to run the intensive computations required by the project (WPs 3 and 4), we will use project funds to acquire at least five GPUs (Graphics Processing Units). Because of their highly parallel structure, the latter are much more efficient than CPUs in implementing algorithms, such as neural networks, involving massively parallel computation. This activity runs through the whole project to address any potential infrastructure issues.

WP2 Activity 2: Release of code and data. (Lead: PI; participants: everybody) [Months 31-60]
This Activity ensures that all the research output is made publicly available in a properly documented form, to enable replication of published results and further research. The code of the model will be posted to public online repositories such as GitHub, and pre-trained models and datasets will be available for download from the project's website (see WP 5) upon publication of the relevant results. The final versions of all of the digital output will also be made available through the university's e-repository, that ensures data preservation.

Milestones and deliverables

M1.1.1	GPUs up and running	M6
M1.1.2-4	Infrastructure review	M24,36,48
M1.2.1-2	Code cleanup and review	M29,53
M1.2.9	Digital data review	M54
D1.1.1	Computational infrastructure documentation	M9
D1.2.1-4	Public pre-trained model and dataset release for WP 4.1, 3.1, 3.2, 4.2	M36,42,60,60
D1.2.5-6	Public code release	M30,54
D1.2.7	Digital output release in university e-repository	M60

Work Package 2: Model development [Months 1-60]

Goals: develop and implement the model; assess its scientific implications.[7]

WP2 Activity 1: Model definition and implementation. (Lead: PI; participants: L. McNally, post-doc 1) [Months 1-54]
We develop the formal definition of the model and its implementation as a neural network. This Activity lasts until month 54 to adjust the model according to the experience gathered in the project and address any issues that may arise. We will start from the proposal in Section a.3, and experiment with alternative architectures. We will also explore general initialization strategies, such as seeding the distributional lexicon (component (A) in Figure 1 above) with high-quality embeddings. For WP 3 (linguistic and visual input), we will use pre-trained multimodal word representations combining linguistic and visual information (Lazaridou et al. 2015) and a pre-trained convolutional neural network (Krizhevsky et al. 2012) to map images to the visual layer of the model. See WP3 and WP4 for other, task-dependent transfer learning strategies.

WP2 Activity 2: Task adaptation (Lead: PI; participants: post-doc 1) [Months 12-54]
We define how the network is adapted to the input and target output required by each task. For instance, for the "Ann" task in WP 3.1, the required output is the name of the character (e.g. "Ann") referred to by a noun phrase (e.g., "The Medicine student"). The network first processes the whole story one word at a time and creates an entity library E^n (where *n* indexes the last word in the story). At query time, the network processes the definite description of interest also one word at a time, with vector $\mathbf{i}_d$ corresponding to the state of the integrated layer (see Fig. 1) after the description has been read. The most related entity in the entity library is then "softly" retrieved by first measuring similarity of all entity vectors to $\mathbf{i}_d$ (Equation (7)) and then using the resulting normalized similarities **s** as weights in a weighted sum of the entity vectors that produces the "relevant" vector **r** (Equation (8)). Finally, normalized similarities of **r** with the vectors representing the candidate names in the distributional lexicon layer (inside component A in Fig. 1) are returned as guesses about the name the target description refers to, via Equation (9), where R embeds **r** in the lexicon layer space, and the W_{names} columns store the candidate name representations.

(7) $\mathbf{s}=(softmax(\mathbf{i}_d^T E^n))^T$

(8) $\mathbf{r} = E^n\mathbf{s}$

(9) $\mathbf{g}=(softmax((R\mathbf{r})^T W_{names})$

[7] This is not assigned a specific Activity because it will be carried out continuously.

During training, the model guesses are compared to the correct response, and the model parameters are consequently adjusted using standard error backpropagation and gradient descent techniques (Nielsen 2015). The architecture for the experiments with images (WP 4) is analogous, except that the input includes image representations presented alone or in parallel with the last word in the corresponding verbal descriptions. In order to identify characters in movie scripts (WP 3.2), we will pre-process the scripts so that character names follow their lines, and we will let the network, at each step, generate probability distributions over the next word through a softmax layer on top of **i**, as in a standard language modeling task. However, in our evaluation we will focus on the ability of the model to produce the character names, rather than all words in the corpus. In this case, the entity library will contribute to the task through the summary vector **b** influencing **i** (see a.3).

Milestones and deliverables

M2.1.1-3	Model defined in iteratively improved versions	M9,30,42
M2.1.4-6	Model implemented in iteratively improved versions	M18,36,48
M2.2.1-2	Model adapted to WP3.1, WP3.2 tasks	M18,36
M2.2.3-4	Model adapted to WP4 tasks (perception only, integrated)	M21,33
D2.1.1-3	Model definition and implementation report	M54
D2.2.4-6	Internal release of code	M19,37,49

Work Package 3: Linguistic knowledge-based reference [Months 7-54]

Goals: test the model on reference to entities based on knowledge gathered from the previous linguistic discourse; assess the ability of the model to handle entity representations for entities that are newly introduced in a discourse and previously known entities.

WP3 Activity 1: Reference to previously-unknown entities (Lead: PI; participants: post-doc 2, PhD student 1) [Months 7-36]

The first experiment concerns reference to previously unknown entities introduced in a discourse: Their dynamic introduction to the entity library as they are mentioned, the update of their representation based on the linguistic information provided the discourse, and their retrieval based on the task. The task we will use for this experiment is the "Ann" task introduced above. For the goals of the experiment (dynamic initialization and characterization of distinct entities), it is important that the entities not be previously known and that several entities are introduced within a limited discourse, so we cannot use already available narrative text: News and Wikipedia articles typically concern entities in the public domain (Obama, England, etc.); novels and even short stories are too long; and there are not enough freely available, short enough narrative units for our purposes. Therefore, we develop our own dataset, as follows. We will first ask human subjects (via crowdsourcing) to create stories using four, randomly selected different proper nouns. The pool of proper nouns will be harvested from publicly available sources such as Wikipedia, with filters to ensure that they do not refer to unique public entities (such as Obama or England). Once we have the stories, we will obtain referring expressions through an online game that resembles the popular Taboo. In the game, two players are randomly paired; the system asks player A to produce a definite description for one of the characters (say, "the Medicine student"); Player B has to guess which of the four characters it refers to (in the example, "Ann"). Player A is not allowed to use more than a limited number of the words that appear in the story, to avoid too literal repetitions of expressions. Note that we are using methodology that has been shown to work before for the construction of related data sets: For the MCTest data set, subjects were asked to write short stories for a more general Machine Comprehension of Text task (Richardson), and the ReferItGame data set mentioned above used a similar game to produce referring expressions for objects in images (Kazemzadeh et al. 2014). Based on these previous experiences, a realistic goal is to gather a dataset containing 160,000 referring expressions for entities in 40,000 stories, which will facilitate empirical approaches to reference and will be very useful for creating further resources for related tasks such as coreference resolution, Machine Comprehension of Text, and Question Answering. For instance, the MCTest consists of only 500 stories and 2,000 reading comprehension questions about those stories. The most resource-consuming part is building the stories, which is handled by our project. It will thus be feasible for other researchers to extend MCTest by producing reading comprehension questions for the stories in our dataset.

We will consider the following existing data sets for related tasks to pre-train the model (transfer learning) and/or extrinsic evaluation: the OntoNotes corpus for coreference resolution (Hovy et al. 2006); the DeepMind data set (Hermann et al. 2015) containing news articles with an associated fill-in-the-blank task about entities; and the Children's Book Test dataset (Hill et al. 2015) providing another fill-in-the-blank task for proper nouns, common nouns, verbs, and prepositions.

WP3 Activity 2: Long-term modeling of entities (Lead: PI; participants: post-doc 2, PhD student 1) [Months 37-54]

The previous experiment tests the construction of newly-introduced entity representations. Here we test the ability of the model to handle previously known entities (recall the "Aunt Lina" example from Section a.3), through a task that thoroughly probes the entity library across time. The task is to associate each utterance in the second part of a movie script with the character that produced it, based on the entity library built in the first part. We operationalize it as a language modeling task, where the output is a probability distribution over the whole vocabulary, and evaluate only on the production of character names. We will give the system the first part of each movie as input (so it constructs the entity representations) and test on the second part (so that it updates it and retrieves information from it). The task has the advantage that we can produce a large amount of training data automatically, by collecting and processing freely available movie scripts from dedicated sites such as the Internet Movie Screenplay Database.[8]

Milestones and deliverables

M3.1.1	Design for story and noun phrase collection ready	M8
M3.1.2	Pilot data set study done	M11
M3.1.3	"Ann" data set finished	M20
M3.1.4	Reference to newly-introduced entities experiments and analyses done	M32
M3.2.1	Data set of movie scripts finished	M36
M3.2.2	Long-term modeling of entities experiments and analyses done	M48
D3.1.1	"Ann" data collection protocol reports	M21
D3.1.2	Internal data set release	M22
D3.1.3	Report for reference to newly-introduced entities experiments	M36
D3.2.1	Movie script dataset creation protocol report	M37
D3.2.2	Internal movie script dataset release	M38
D3.2.3	Report for long-term modeling of entities experiments	M54

Work Package 4: Perception-based and integrated reference [Months 7-54]

Goal: test the model on reference using perceptual (visual) information, on its own or integrated with previous knowledge acquired through language.

The experiments in this WP concern the identification of an entity depicted in an image through a noun phrase with descriptive content (e.g., using "the white cat" to identify a white cat in a picture). Because this is an ambitious task that requires the integration of all the components in the model, we carry out the experiments in two phases with a more controlled but less natural setup (Activity 1) and a more natural but also computationally more difficult setup (Activity 2), respectively.

WP4 Activity 1: Controlled domain (Lead: PI; participants: post-doc 3, PhD student 2) [Months 7-30]

We test reference to entities based on visual properties only (experiment 1) and visual properties plus knowledge gained through language (experiment 2). In experiment 1, we will present the system with a sequence of images (for instance, a plant, a white cat, a black dog, a black cat, a white dog) followed by one noun phrase. The task of the system is to decide whether it has seen a unique entity corresponding to that noun phrase. In the example, the system should reply "yes" to "white cat", "no" to "white animal" (because it has seen two), and "no" to "brown dog". It is important that the images be associated not only with the general object category ("cat") but also with properties ("white"), such that we can include entities of the same category with different properties; this should direct the model to learn entity representations, as opposed to object categories. Note that since multiple images (objects) are presented, and the model is tested after their sequential presentation, this and the following tasks are crucially probing the models' ability to store representations of different entities in its library, keeping them distinct through time.

Both the visual and the linguistic data used in this experiment are from a controlled domain, to maximize the feasibility of the task and the inspectability of the model. We will use a pre-existing data set (Russakowski and Li 2012) with 10,000 images of single objects annotated with visual attributes. The data set contains 25 images for each of roughly 400 different object categories (*cat*, *plant*, etc.). Each image is exhaustively annotated with 25 visual attributes (*white*, *furry*, *metallic*, etc.). We will create noun phrases describing each of the images from the annotation in the dataset (from attribute *white* and object category *cat* to the noun phrase "the white cat"). Some linguistic variation in the noun phrases will be achieved by automatic procedures, such as using hypernyms from an existing taxonomy ("animal" for an image of a cat; we will use WordNet, Miller 1995). We will then construct sequences of five images depicting different objects, from dissimilar to same-category (again relying on WordNet). We will reuse the images in different

[8] http://www.imsdb.com.

sequences and test different noun phrases, thus creating enough datapoints to train the model. If necessary, however, we could pre-train on noisy data and do transfer learning (see below).

Experiment 2 integrates two sources of information about entities: Perceptual (visual) information, and previous knowledge coming from language. We will add further information about the entities, associating each image in the input with linguistic expressions (e.g., "is sociable" for the white cat image), and evaluate the model on noun phrases using only visual information ("white cat"), only language-based knowledge ("is sociable"), and an integration of the two ("sociable white cat", "sociable animal"). The linguistic expressions will be harvested from available textual corpora, ensuring that they are at the same time consistent with the object depicted in the image (e.g., that we find instances of sentences like "my cat is sociable") and not redundant (redundancy will be estimated using the distance in semantic space between the linguistic expression, e.g. "is sociable", and the noun, e.g. "cat"). Highly visual linguistic expressions like "plump" will be avoided using a filter based on available concreteness scores for words (Turney et al. 2011), to prevent a potential clash in the visual and linguistic inputs, and to ensure that visual and linguistic information are complementary.

WP4 Activity 2: Open domain (Lead: PI; participants: post-doc 3, PhD student 2) [Months 31-54]

Actvity 1 uses a very restricted setup: few attributes, automatically constructed referring expressions, and images depicting a single object. This Activity carries out experiments analogous to those in Activity 1, but using naturalistic images with human-produced referring expressions. One technical difference between the previous version of the tasks and the current ones is that the system will get only one image as input, because the images depict multiple objects. We will use the ReferItGame dataset (Kazemzadeh et al. 2014), which contains 130,000 human-generated referring expressions for the objects in 20,000 natural images where the objects belong to 238 categories. The boundaries of the objects in each image have been previously identified. We will use the same methodology as in Activity 1 to create the linguistic expressions that we will associate with each object in the input image in the experiment, testing integrated reference. We will also experiment with pre-training the model with data automatically extracted from existing caption datasets (Young et al. 2014; Chen et al. 2015), which are very large but do not come with pre-identified object boundaries nor alignment between the referring expressions in the captions and the objects.

Milestones and deliverables

M4.1.1	Datasets for experiments in the controlled domain created	M9,24
M4.1.2	Experiments and analyses for the controlled domain done	M21,30
M4.2.1	Datasets for experiments in the open domain created	M36,48
M4.2.2	Experiments and analyses for the open domain done	M45,54
D4.1.1	Controlled domain data set creation protocol reports	M10,25
D4.1.2	Internal release of controlled domain datasets	M11,26
D4.1.3	Report for controlled domain experiments	M30
D4.2.1	Open domain data set creation protocol reports	M37,49
D4.2.2	Internal release of open domain datasets	M38,50
D4.2.3	Report for open domain experiments	M54

Work Package 5: Management and dissemination [Months 1-60]

Goals: ensure good project development through planning, coordination, and dissemination activities; assess progress and readjust targets.

WP 5 Activity 1: Project management (Lead: PI; participants: everybody) [Months 1-60]

I will ensure that the project runs smoothly and communication flows, and will lead planning and coordination, with the support of the rest of the team. Weekly project meetings will discuss both organizational and scientific matters, and meetings with subsets of the team will be held as needed. A reading group will meet weekly or biweekly through the duration of the project to create a shared body of knowledge and keep up with the relevant literature, and internal tutorials regarding formal and computational content (including toolkits and resources) will be organized on demand. The three External Advisory Board members will visit after months M18, M36, and M54 to assess progress, provide feedback, and identify future research and funding opportunities. I will also oversee financial and administrative management, including ethical issues, with the support of the research service at Universitat Pompeu Fabra.

WP 5 Activity 2: Dissemination (Lead: PI; participants: everybody) [Months 1-60]

I and the rest of the AMORE team will ensure adequate dissemination of project results. A simple but informative project website will be set up in the beginning of the project, providing updated information and pointers to publications, code, and data. We will also use the online academic social networks of all the team members (e.g. ResearchGate) and the university to publish updates about the project.

Project results will be disseminated to the research community through publications and talks at international scientific events. Since Computational Linguistics is a heavily conference-oriented field, proceeding publications will be the main instrument for periodic updates about the project (candidate venues include ACL, NAACL, EACL, EMNLP, IWCS, and *SEM; as well as SALT and *Sinn und Bedeutung* for theoretical semantic results). Its main results will be published in journals: We will publish at least one general Computational Linguistics article on WP 4.1, one article in a journal focusing on linguistic resources about the data set in WP 4.1, one Artificial Intelligence journal article on WP 5, and possibly one theoretical linguistics journal article on the implications of our project for semantic theory and an overall summary in a high impact general science journal. Also, we will organize two workshops to amplify the results of the project and foster research on linguistic reference to entities: One in the middle of the project, co-located with an international conference, and one at the end of the project with invited talks by prominent researchers in the field, to summarize and appraise the project results and discuss future research.

We will also disseminate our research to society through outreach activities, participating in the European Researchers' Night and the EscoLab programme of the Barcelona City Council to promote science directed at secondary school students. Candidate activities include a demonstration of the reference game via which we will collect our data set and an interactive demonstration of the system on tasks WP 3.1 and WP 4.1.

Milestones and deliverables

M5.1.1	GPUs purchased	M3
M5.1.2	External personnel (excluding students) hired	M6
M5.1.3	Review by the university's Ethical Committee obtained	M6
M5.1.4	PhD students selected	M12
M5.1.5-7	External consultant reviews	M18,36,54
M5.2.1,2	Project website up; all project deliverables available from website	M3,60
M5.2.3,4	Mid-project and final workshops	M36,54
M5.2.5,6	European Researchers' Night and EscoLab outreach activities	M36,48
D5.1.1-3	Review reports	M19,37,55
D5.1.4,5	PhD theses defended	M60
D5.2.1	5 conference articles on specific modeling experiments, dataset construction, and linguistic issues (dates submitted)	M26,27,36,38,51
D5.2.2	5 journal articles: about the data set in WP 3.1, the model and experiments in WP 3, the model and experiments in WP 4, implications for semantic theory, and general scientific significance (dates submitted)	M36,39,54,60,60

Section c. Resources (including project costs)

c1. Personnel involvement timeline.

Personnel involvement timeline

	M6	M12	M18	M24	M30	M36	M42	M48	M54	M60
Boleda (PI)										
McNally										
Post-doc 1										
Post-doc 2										
Post-doc 3										
PhD student 1										
PhD student										

Senior staff:

Gemma Boleda (PI, 75% dedication for 5 years) and Louise McNally (10% dedication for 5 years), an expert in formal semantics and long-term collaborator of mine who will participate in the formal specification of the model (WP2) and in assessing the implications of the project for theoretical linguistics.

Personnel to be hired with project funds:

Post-doc 1 (advanced expertise in neural networks) will develop and implement the model (WP2). **Post-doc 2** (expertise in distributional semantics and/or discourse) and **PhD student 1** will carry out the experiments on linguistic knowledge-based reference (WP3). **Post-doc 3** (expertise in integrating language and vision) and **PhD student 2** will carry out the experiments on perception-based and integrated reference (WP4). The topics of the PhD theses will be modeling linguistic knowledge-based reference (PhD student 1) and reference integrating perceptual information (PhD student 2).

External consultants:

- Katrin Erk (The University of Texas at Austin, USA): expert on both formal and distributional semantics.
- Hans Kamp (University of Stuttgart, Germany): expert on formal semantics, symbolic computational linguistic, and philosophy of language; creator of Discourse Representation Theory and the model of reference that inspires ours.
- Hinrich Schütze (University of Munich, Germany): expert on distributional semantics and Machine Learning, including neural networks.

c2. Summary of costs

See c.1 for personnel. Travel covers consultant visits and two conferences per year per member (including registration fees, which for Computational Linguistic conferences are around 500€). Equipment includes GPU units (see WP1.1; 35,000€ in total), one computer per postdoc (as the department does not supply them), and books. Publication costs are expected to be low because some target journals do not charge publication fees. *Other* covers the construction of the dataset in WP 3.1 (140,000€) and the organization of the two workshops (see WP 5.2), including the travel expenses of the invited speakers.

Cost Category			**Total in Euro**
Direct Costs	**Personnel**	PI	158291
		Senior Staff	36502
		Postdocs	528991
		Students	165560
		Other	0
	i. Total Direct Costs for Personnel (in Euro)		889344
	Travel		98000
	Equipment		40500
	Other goods and services	Consumables	0
		Publications (including Open Access fees), etc.	7000
		Other (please specify)	165000
	ii. Total Other Direct Costs (in Euro)		310500
A – Total Direct Costs (i + ii) (in Euro)			1199844
B – Indirect Costs (overheads) 25% of Direct Costs (in Euro)			299961
C1 – Subcontracting Costs (no overheads) (in Euro)			0
C2 – Other Direct Costs with no overheads (in Euro)			0
Total Estimated Eligible Costs (A + B + C) (in Euro)			1499805
Total Requested EU Contribution (in Euro)			1499805

Please indicate the duration of the project in months:	**60**
Please indicate the % of working time the PI dedicates to the project over the period of the grant:	**75%**
Please indicate the % of working time the PI spends in an EU Member State or Associated Country over the period of the grant:	**100%**

I am fully committed to AMORE for the whole duration of the project, except for about 80 hours of teaching per year, administrative duties related to my position, and student supervision (one PhD student outside the project, working on related topics).

REFERENCES

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). VQA: Visual

Question Answering. *arXiv preprint arXiv:1505.00468*.
Arsenijević, B., Boleda Torrent, G., Gehrke, B., & McNally, L. (2010). Ethnic adjectives are proper adjectives. In Baglini, R., Grinsell, T., Keane, J.A., Singerman, R. & Thomas, J. (Eds.), *CLS 46-I The Main Session: Proceedings of 46th Annual Meeting of the Chicago Linguistic Society*, 17–30.
Asher, N. (2011). *Lexical Meaning in Context: A Web of Words*. Cambridge University Press.
Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of ICLR*, 1–15.
Baroni, M., & Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, *36*(4), 673–721.
Baroni, M., & Zamparelli, R. (2010). Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of EMNLP*, 1183–1193.
Baroni, M., Bernardi, R., & Zamparelli, R. (2014a). Frege in space: A program of compositional distributional semantics. *Linguistic Issues in Language Technology*, *9*.
Baroni, M., Dinu, G., & Kruszewski, G. (2014b). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of ACL*, 238–247.
Boleda; G. & Herbelot, A. (in prep.). Special Issue on Formal Distributional Semantics. *Computational Linguistics*. In preparation, to be published in the fall of 2016.
Boleda, G., im Walde, S. S., & Badia, T. (2007). Modelling Polysemy in Adjective Classes by Multi-Label Classification. In *Proceedings of EMNLP-CoNLL*, 171–180.
Boleda, G., Evert, S., Gehrke, B., & McNally, L. (2012a). Adjectives as saturators vs. modifiers: Statistical evidence. In Aloni, M., Kimmelman, V., Roelofsen, F., Sassoon, G. W., Schulz, K. & Westera M. (Eds.) *Logic, Language and Meaning - 18th Amsterdam Colloquium, Amsterdam, The Netherlands, December 19-21, 2011, Revised Selected Papers*, 112–121. Springer.
Boleda, G., Padó, S., & Utt, J. (2012b). Regular polysemy: A distributional model. In *Proceedings of *SEM*, 151–160.
Boleda, G., Vecchi, E. M., Cornudella, M., & McNally, L. (2012c). First-order vs. higher-order modification in distributional semantics. In *Proceedings of the EMNLP-CONLL*, 1223–1233.
Boleda, G., im Walde, S. S., & Badia, T. (2012d). Modeling regular polysemy: A study on the semantic classification of Catalan adjectives. *Computational Linguistics*, *38*(3), 575–616.
Boleda, G., Baroni, M., The Pham, N. & McNally, L. (2013). Intensionality was only alleged: On adjective-noun composition in distributional semantics. In *Proceedings of IWCS*, 35–46.
Boleda, G., & Erk, K. (2015). Distributional Semantic Features as Semantic Primitives—Or Not. In *Knowledge Representation and Reasoning: Integrating Symbolic and Neural Approaches. Papers from the AAAI Spring Symposium*.
Beltagy, I., Chau, C., Boleda, G., Garrette, D., Erk, K., & Mooney, R. (2013). Montague meets Markov: Deep semantics with probabilistic logical form. *Proceedings of* SEM*, 11–21.
Borges, J. L. (1944). Funes El Memorioso. In *Ficciones (1935-1944)*. Ediciones SUR. Citing from the translation by J. E. Irby, 2000, Funes the Memorious.
Bos, J., & Markert, K. (2005). Recognising textual entailment with logical inference. In *Proceedings of HLT-EMNLP*, 628–635.
Bos, J. (2008). Wide-coverage semantic analysis with Boxer. In Johan Bos and Rodolfo Delmonte (eds.), *Semantics in Text Processing*, 277–286. College Publications.
Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of EMNLP*, 632–642.
Bruni, E., Boleda, G., Baroni, M., & Tran, N. K. (2012). Distributional Semantics in Technicolor. In *Proceedings of ACL*, 136–145.
Chen, X., Fang, H., Lin, T.-Y., Gupta, R. V. S., Dollár, P., & Lawrence Zitnick. (2015). Microsoft COCO Captions: Data Collection and Evaluation Server. *arXiv preprint arXiv*: 1504.0032.
Chomsky, N. (1957). *Syntactic structures*. Walter de Gruyter.
Churchland, P. M. (1998). Conceptual similarity across sensory and neural diversity: The Fodor/Lepore challenge answered. *The Journal of Philosophy*, 5–32.
Croft, W. & Cruse, A. (2004). *Cognitive Linguistics*. Cambridge University Press.
Cruse, D. A. (1986). *Lexical semantics*. Cambridge University Press.
van Deemter, K. (2010). *Not exactly: In praise of vagueness*. Oxford University Press.
Elman, J. L. (1990). Finding structure in time. *Cognitive science*, *14*(2), 179–211.
Firth, J. R. (1957). *Papers in Linguistics 1934–1951*. Oxford University Press.
Fodor, J. A., Garrett, M. F., Walker, E. C., & Parkes, C. H. (1980). Against definitions. *Cognition*, *8*(3), 263–367.

Fodor, J., & Pylyshyn, Z. (1988). Connectionism and Cognitive Architecture: a Critical Analysis. *Cognition*, 28: 3–71.

Fodor, J., & Lepore, E. (1999). All at sea in semantic space: Churchland on meaning similarity. *the Journal of Philosophy*, 381–403.

Frege, G. (1892). Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, *100*, 25–50.

Gupta, A., Boleda, G., Baroni, M., & Padó, S. (2015). Distributional vectors encode referential attributes. In *Proceedings of EMNLP*, 12–21.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1), 335–346.

Harris, Z. (1968). *Mathematical Structures of Language*. Wiley.

Herbelot, A., & Vecchi, E. M. Building a shared world: Mapping distributional to model-theoretic semantic spaces. In *Proceedings of EMNLP*, 22–32.

Hermann, K. M., Kočiský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. (2015). Teaching Machines to Read and Comprehend. *arXiv preprint arXiv:1506.03340v1*.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., Senior, A. ; Vanhoucke, V. ; Nguyen, P. ; Sainath, T.N. ; Kingsbury, B. & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine, 29*(6), 82–97.

Hill, F., Bordes, A., Chopra, S., & Weston, J. (2015). The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations, 1–13. *arXiv preprint arXiv:*1511.02301.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*8), 1–32.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., & Weischedel, R. (2006). OntoNotes: the 90% solution. In *Proceedings of NAACL*, 57–60.

Ji, Y., & Eisenstein, J. (2015). One Vector is Not Enough : Entity-Augmented Distributed Semantics for Discourse Relations. *Transactions of the ACL*, *3*, 329–344.

Joulin, A., & Mikolov, T. (2015). Inferring Algorithmic Patterns with Stack-Augmented Recurrent Nets. *arXiv preprint arXiv:1503.01007*.

Kalchbrenner, N., & Blunsom, P. (2013). Recurrent Convolutional Neural Networks for Discourse Compositionality. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, 119–126.

Kamp, H. (1981). A theory of Truth and Semantic Representation. In Groenendijk, J., Janssen, T. H. & Stokhof, M. (Eds.) *Formal Methods in the Study of Language, Part I*. Mathematisch Centrum, 277-322.

Kamp, H. (2015). Entity Representations and Articulated Contexts An Exploration of the Semantics and Pragmatics of Definite Noun Phrases. Ms., University of Stuttgart.

Kamp, H., & Reyle, U. (1993). *From discourse to logic*. Kluwer.

Katz, J. J., & Fodor, J. (1963). The Structure of a Semantic Theory. *Language*, *39*(2), 170–210.

Kazemzadeh, S., Ordonez, V., Matten, M., & Berg, T. L. (2014). ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of EMNLP*, 787–798.

Keefe, R. (2000). *Theories of vagueness*. Cambridge University Press.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

Kruszewski, G. & M. Baroni (2015). So similar and yet incompatible: Toward automated identification of semantically compatible words. In *Proceedings of NAACL HLT*, 964–969.

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

Le, P., & Zuidema, W. (2014). The inside-outside recursive neural network model for dependency parsing. In *Proceedings of EMNLP*, 729–739.

LeCun, Y., Bengio, Y. & Hinton, G. E. (2015) Deep Learning. *Nature*, *521*, 436–444.

Lewis, M., & Steedman, M. (2013). Combining Distributional and Logical Semantics. *Transactions of the Association for Computational Linguistics*, *1*, 179–192.

Li, J., Jurafsky, D., & Hovy, E. (2015). When Are Tree Structures Necessary for Deep Learning of Representations? In *Proceedings of EMNLP*, 2304–2314.

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9(85)*, 2579–2605.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT press.

Marelli, M., Bentivogli, L., Baroni, M., Bernardi, R., Menini, S., & Zamparelli, R. (2014). SemEval-2014 Task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of SemEval*, 1–8.

Margolis, E. & S. Laurence (1999). *Concepts: Core Readings*. MIT Press.
Mayol, L., G. Boleda & Badia, T. (2005). Automatic acquisition of syntactic verb classes with basic resources. *Language Resources and Evaluation*, 39(4), 295–312.
McNally, L. & Boleda, G. (2004). Relational adjectives as properties of kinds. In Olivier Bonami and Patricia Cabredo Hofherr (eds.) *Empirical Issues in Syntax and Semantics 5*, 179–196.
McNally, L. & Boleda, G. (2015). Conceptual vs. Referential Affordance in Concept Composition. To appear in Winter, Y. & Hampton, Y. (eds.) *Concept Composition and Experimental Semantics/Pragmatics*. Springer.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.
Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39–41.
Mitchell, J., & Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science*, *34*(8), 1388–429.
Montague, R. (1970). English as a formal language. In Bruno Visentini (Ed.) *Linguaggi nella società e nella tecnica*. Mailand 1970, 189–223. Reprinted in (Thomason 1974), 188–221.
Montague, R. (1960). On the nature of certain philosophical entities, *The Monist*, 53: 159–194. Reprinted in (Thomason 1974), 148–187.
Murphy, G. L. (2002). *The big book of concepts*. MIT press.
Nielsen, M. A. (2015). *Neural Networks and Deep Learning*. Determination Press.
Poesio, M., Stuckardt, R. & Versley, Y. (in press). *Anaphora Resolution: Algorithms, Resources, and Applications*. Springer.
Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge, MA (etc.): The MIT Press.
Richardson, M., Burges, C. J. C., & Renshaw, E. (2013). MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text. In *Proceedings of EMNLP*, 193–203.
Rocktäschel, T., Grefenstette, E., Hermann, K. M., & Blunsom, P. (2015). Reasoning about Entailment with Neural Attention. *arXiv preprint arXiv:* 1509.06664v1
Roller, S., Erk, K., & Boleda, G. (2014). Inclusive yet selective: Supervised distributional hypernymy detection. In *Proceedings of COLING*, 1025–1036.
Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review, 65(6), 386.
Russakovsky, O., & Fei-Fei, L. (2012). Attribute learning in large-scale datasets. In *Trends and Topics in Computer Vision*, 1–14. Springer.
Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, *3*(03), 417–424.
Smolensky, P. (1987). The Constituent Structure of Connectionist Mental States: A Reply to Fodor and Pylyshyn. *The Southern Journal of Philosophy*, 26 (Supplement): 137–161.
Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of EMNLP*, 1631–1642.
Sukhbaatar, S., Szlam, A., Weston, J., & Fergus, R. (2015). End-To-End Memory Networks, 1–11. *arXiv preprint arXiv*:1503.08895
Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, *37*(1), 141–188.
Turney, P. D., Neuman, Y., Assaf, D., & Cohen, Y. (2011). Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of EMNLP*, 680–690.
Thomason, R. H. (Ed.) (1974). *Formal Philosophy, Selected Papers of Richard Montague*. Yale University Press.
Vallduví, E., & Engdahl, E. (1996). The linguistic realization of information packaging. *Linguistics*, *34*(3), 459–520.
Voorhees, E. M. (1999). The TREC-8 Question Answering Track Report. In *Proceedings of TREC*, 77–82.
Weiss, D., Alberti, C., Collins, M., & Petrov, S. (2015). Structured training for neural network transition-based parsing. *arXiv preprint arXiv*:1506.06158.
Werbos, P. J. (1988). Generalization of backpropagation with application to a recurrent gas market model. *Neural Networks*, *1*(4), 339–356.
Weston, J., Chopra, S., & Bordes, A. (2014). Memory networks. *arXiv preprint arXiv:1410.3916*.
Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.
Wittgenstein, L. (1953). *Philosophical Investigations*. G.E.M. Anscombe and R. Rhees (eds.), G.E.M.

Anscombe (trans.), Oxford: Blackwell.
Young, M. H. P., Lai, A., & Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, *2*, 67–78.