

ACQUIRING AND REPRESENTING MEANING: THEORETICAL AND COMPUTATIONAL PERSPECTIVES

Alessandro Lenci

Università di Pisa – Dipartimento di Linguistica “T. Bolelli” (Pisa, Italy)
alessandro.lenci@ilc.cnr.it

Simonetta Montemagni

Istituto di Linguistica Computazionale – CNR (Pisa, Italy)
simonetta.montemagni@ilc.cnr.it

Vito Pirrelli

Istituto di Linguistica Computazionale – CNR (Pisa, Italy)
vito.pirrelli@ilc.cnr.it

1. Introduction

Modelling the way word meanings dynamically function and combine in communicative contexts, evolve through learning and are categorised through high-level semantic classes presents one of the most difficult challenges for cognitive science, and is a large stumbling block on the way to developing advanced artificial systems for full text understanding. The problem of the form of semantic knowledge typically presents itself as a *representation* issue: i.e. what is the aptest way of representing the meaning of words and of the complex expressions they enter into? Providing satisfactory answers to these questions is an essential requirement for explaining the effective use of semantic knowledge in concrete cognitive abilities. This is extremely important also in an engineering

and computational perspective, as a key to a deeper understanding of the constructive principles underpinning the design of intelligent artifacts like robots and other artificial intelligent agents. Similarly, the issue of *semantic dynamics* has a crucial role in modelling human cognition, since cognitive agents constantly update and revise their knowledge, acquire new words, assign new meanings to already known words, etc. Word meanings also exhibit a protean nature, a multifaceted behaviour closer to a kaleidoscope of senses continuously changing their relations and nature depending on the perspective from which they are observed. On the more application-oriented side, endowing systems with the capability of extending and adapting their semantic knowledge is an essential precondition to robustness, scalability, and effectiveness in tackling real-world tasks.

One of the main tenets of contemporary research into the structure and function of language is that *understanding core lexical semantic properties amounts to providing an explicit, formal way to represent word meaning*. In particular, semantic representations are purported to play a key role in explaining the following central phenomena:

1. *multiple meanings* – in most cases the information content of a word changes dramatically depending on the linguistic context in which it appears. This is traditionally referred to as the fact that a word may have more than one *sense*. A classical philosophical and linguistic conundrum is how the existing partitions within the semantic space of a word can be captured in terms of lexical representations. This in turn requires that we can characterize and distinguish the different meanings of a word, set the boundaries and relations between them, and link these meanings to the various types of contexts that select them. This is related to the well-known problems of *lexical ambiguity*, *polysemy*, *word-sense disambiguation*, etc. A further issue concerns whether and to which extent meaning multiplicity belongs to the stable long term organization of lexical representations, or is instead generated “on line”, and in the latter case with the help of which generative devices (cf. Pustejovsky 1995);
2. *lexical inferences* – understanding the meaning of a word in a certain context also means to be able to draw a number of inferences. This inferential competence (Marconi 1997) is a crucial part of

our knowledge of the meaning of a word. For instance, from *John bought a bottle of wine*, we can easily infer that John bought a bottle containing a drinkable and tasty liquid substance. Moreover, from *Mary bought a car from Ted*, it is possible to infer that Mary has now a car previously belonging to Ted. The question is here to explain the *prima facie* indisputable fact that inferences like these are deeply related to what *wine* and *buy* mean in the context of those sentences. The open issue is whether this intuition can be grounded to the actual organization of lexical representations, and – in the case that a distinction between lexicon and context driven inferences is really possible – how to establish the boundary between the two realms;

3. *semantic similarity judgments* – when we understand the meaning of a word we are also able to decide how similar it is either to the way other words are used, or to other usages of the same word. For instance, the meaning of the verb *open* in *John opened the door* is strikingly more similar to its meaning in *The door opened* rather than to the one in *John opened the meeting*. This seems to be true notwithstanding the fact that in the first and third sentence, but not in the second one, *open* occurs transitively and with an agentive interpretation. However, if the latter aspect is focused on, our similarity judgments can well be reversed. Actually, similarity judgments are claimed to have a central and widespread role in cognition. Usually, we group together similar things, and this is the main source of *categorization* and *concept formation* processes. In the same way, words with similar meanings appear to form semantic paradigmatic classes that tend to distribute with similar structures. In fact, one of the major problems with similarity judgments is their being necessarily dependent on the context and on the perspectives under which we express them (cf. Barsalou 1982, Medin *et al.* 1993). This makes explaining semantic similarity as a relation between lexical representations a highly challenging and often problematic endeavor, in turn critically dependent on the nature itself of semantic representations. As Goodman (1972) claims, similarity has an insidious nature, so that it is able to defy every theory attempting to grasp and define it;

4. *meaning composition* – the information content of a word constrains its possibilities to combine with other words, and also the result of such combinations. A core part of lexical semantic representations should therefore concern the specification of the possible semantic contexts in which it can or must appear. Well-known notions such as *argument structure*, *selectional restrictions* and *thematic relations* are intended to address these properties of meaning representations. The key challenge however remains how to identify those aspects of the meaning of a word that are responsible for its selective and distributional potential. Actually, there is little sense in inquiring lexical meaning without referring to the *syntagmatic processes* it participates in. The regularities between lexical semantic properties and syntactic structures are often regarded as playing a key role in language acquisition (cf. Pinker 1989). Besides, the linguistic evidence provided by Levin (1993) shows that semantic properties of verbs help predict which syntactic alternations they can take. Lexical representations should therefore provide a suitable explanatory space for the interplay of syntactic and semantic properties.

A central problem in linguistics and cognitive science concerns the form and organization of semantic representations. According to one view (cf. section 2), word meaning representations are to be conceived of as highly articulated, basically static and explicit *symbolic structures* shaping lexical content. Functional processes operate on them on a formal and algebraic basis. At the opposite extreme (cf. sections 4 and 5), we can find approaches that model lexical representations in terms of inherently dynamic, *usage-based word distribution patterns*. In what follows we shall consider the two approaches in some detail, with a view to investigating and assessing their strengths and weaknesses. We shall also entertain and discuss the common sense suggestion that there is something to be gained from their integration and hybridation (cf. section 6).

2. Representing meaning through symbolic structures

Consistently with the cognitive paradigm that regards mind and language as symbolic combinatorial systems, a dominant approach to word meaning representations is in terms of *symbolic structures*. Lexical com-

petence is thus modelled as a *structured formal system of conceptual symbols* onto which lexical terms are projected. Using a term nowadays very popular in Knowledge Representation and computer science, such a conceptual symbol system may be regarded as an *ontology* (cf. Saint-Dizier and Viegas 1995, Vossen 2003). Generally speaking, an ontology defines the set of concepts relevant to the description and organization of a certain domain of knowledge, together with the set of relations and axioms that define its architecture (Gruber 1993, Guarino 1998). Similarly, we can conceive of lexical modelling as the task of managing semantic knowledge in terms of a suitable repertoire of *discrete* semantic types or categories, each corresponding to a certain symbol. To emphasize the specific problems raised by ontologies for word meaning representations, we will refer to them as *lexical ontologies*. Different senses of the same word thus correspond to different elements of the ontology, while its architecture provides an explicit representation of the organization of the lexical space, and an account of the way lexical meanings interact with and relate to each other. Lexical ontologies may wildly vary not only in the repertoire of semantic categories, but also and especially in the type of *structure* assigned to lexical conceptual symbols and in the *architecture* of the overall system.

2.1 Structure as a network

According to one major approach, the structure of the lexical semantic space can be represented as a network of relations connecting word meanings. The information content of a lexical item is then regarded as a point fully identified by its position in the general *semantic network*. The best known example of a network-based lexical ontology, and one of the most striking success stories in computational lexical semantics, is WordNet (Fellbaum 1998), whose main architectural principles have been adopted by a whole family of computational lexicons. Partly sharing the structure of a dictionary and the structure of a thesaurus, WordNet represents the information content of a word w as a number of concepts or senses, each intended as a distinct, atomic semantic object. In turn, a lexical concept C_w is represented as a *synset*, a set including the word w itself and possibly other synonymous terms. Thus, the minimal semantic information unit of the system is formed by the conceptual content shared by synonymous lexical items (with synonymy loosely defined as a certain degree of substitutability in context).

In a network-based lexical ontology, the expressive power of the system in capturing semantic properties is determined by the type and number of lexico-semantic relations interlinking the different points of the conceptual network. In WordNet nouns are organized as *hierarchies* or *taxonomies* of synsets linked by the *hyperonymy/hyponymy* (also known as *subsumption* or *ISA*) relation; verbs and adjectives are instead organized along multiple axes, with hyperonymy having just a secondary role with respect to other relations, such as antonymy, troponymy, entailment, etc. This architecture is based on the view that the hierarchical organization of nominal concepts is a necessary feature of the mental lexicon (Miller 1998: 33), with the subsumption relation acting as the backbone for this particular region of the lexical space. The prominent role assigned to hyperonymy can in fact be explained and justified on various grounds. First of all, giving the hyponym of a word is fairly straightforward and, since Aristotle, looks like the most direct and natural way to define the meaning of the word in question. This also correlates with the traditional lexicographic praxis, according to which senses are defined in terms of their *genus* and *differentia*. Besides, hyperonymy is easily definable in terms of class inclusion, and the identification of the hyperonym/hyponym of a concept is apparently supported by a set of *prima facie* reliable linguistic tests (e.g. *X* is a hyponym of *Y* if the acceptability of *That is a X* entails the acceptability of *That is a Y*, or similarly if *X is Y* is true). Secondly, the taxonomical structure is strictly related to the notion of *property inheritance*, which has a key role in Knowledge Representation. The most common and natural way to organize whatever domain of knowledge seems to be in terms of classes of similar objects forming an *inheritance hierarchy*, in which a subclass inherits the properties of its superordinate. Standard, non-linguistic ontologies used in Knowledge Representation are in fact structured essentially as taxonomies of classes, with transversal relations mostly playing a secondary or accessory role. Last but not least, taxonomies are commonly regarded as being highly salient and central in the organization of adults concepts. Grouping similar things into a hierarchical system where concepts are differentiated into levels of varying specificity and related by class inclusion has a central role in human learning and categorization.

The hypothesis that the nominal lexicon is essentially hierarchically structured is however neither unquestioned, nor without problems from both a linguistic and a cognitive point of view. The predominance of

the taxonomic structure has often been argued to end up collapsing together lexical dimensions that are instead orthogonal. For instance, the acceptability of both *A musician is a performer* and *A dog is a mammal* would allow us to establish a hyperonymy/hyponymy relation between the synset/concept of *musician* and the one of *performer*, as well as between the synset/concept of *dog* and the one of *mammal*. However, the conceptual relation holding between the first two synsets/concepts is strikingly different from the one linking the latter two. A musician is in fact a kind of performer with respect to its particular function or activity, which then in this case acts as the main classificatory dimension. This contrasts with the case of natural categories (such as *cat* or *mammal*) in which properties like shape, constitution, colour, etc. are typically the most salient ones for categorization. This also suggests that linguistic tests as the ones reported above actually underdetermine the real type of conceptual relations that link two concepts. The consequence is therefore that network-based lexical ontologies – WordNet included – risk to suffer from what Guarino (1998) properly calls *ISA overloading*, i.e. the fact that taxonomies actually contain orthogonal and very different lexical dimensions that are nevertheless wholly disguised as *ISA* links. In order to reduce such problems and enhance the expressive power of relational models, some lexical ontologies have greatly extended the array of relations used to represent the noun semantic content, covering various conceptual dimensions other than hyperonymy (e.g. typical function, constitution, etc.). Interesting examples are represented by EuroWordNet (Vossen *et al.* 1998) and SIMPLE (Lenci *et al.* 2000).

The problem raised by the role of hyperonymy in lexicon modelling does not amount to a mere contingent lexicographic issue, but poses genuine and fundamental questions concerning the structure of lexical representations. Busa *et al.* (2001) argue that there is a large group of nouns which are highly resistant to a taxonomic organization. For nouns such as *target* or *priority* asking which proposition of the form *X is Y* is acceptable is almost impossible or at best uninformative. The same holds for most abstract and relational nouns, as it is also proved by the fact that these types of nouns are classically hard cases for whatever naïve lexicographic representation. However, problems like these are not only confined to abstract categories, since, as Jackendoff (2002: 345) also correctly points out, even a concrete noun like *puddle* doesn't really fit into *bodies of water* along with *lakes* and *rivers*. From the cognitive

point of view, Lin and Murphy (2001) offer experimental evidence that non-taxonomic dimensions have an important role in the organization of nominal concepts. While taxonomies have always been regarded as predominant in adults cognition, Lin and Murphy focus on what they call *thematic relations*, including spatial dimensions, functionality, causation, etc. These type of relations are orthogonal to taxonomies, and usually link entities belonging to very different kinds, e.g. an object and its natural location, an event and its participants, etc. The results of Lin and Murphy's experiments suggest a more complex and multidimensional organization of human nominal concepts, in which taxonomical organization is not the only dimension of organization, and in many cases surely not the dominant one. In fact, they argue that taxonomic sorting is not a necessary condition of adult conceptual structure and that many people prefer to sort thematically, even though they categorize and name objects perfectly normally (Lin and Murphy 2001: 6). Accordingly, major architectural differences in the lexicon deeply involve the internal organization of the nominal space itself, in which highly orthogonal principles of organization coexist.

2.2 The functional structure of concepts

One consequence of regarding word meanings as basically atomic entities is that most of the semantic combinatorial properties of lexical items are not explicitly represented in the lexicon. For instance, although synset hierarchies are widely used to express semantic restrictions on predicate arguments, WordNet does not encode any syntagmatic constraint between word senses, such as the number and types of arguments a given lexical item can combine with.

Since Frege, it is common to model the combinatorial properties of lexical items in terms of the *function* and *argument* opposition. A major divide is thus established between predicative lexical items (prototypically verbs, but also relational and event nouns, adjectives, etc.) that project an *argument structure* and those that act as semantic fillers of predicate arguments. In this case, lexical ontologies include conceptual symbols with typed variables, each representing a semantic argument with a certain *thematic role*. The type of the variable is also a conceptual symbol and specifies its *selectional restrictions*, while *thematic roles* (e.g. AGENT, THEME, EXPERIENCER, etc.) label the arguments in terms of their role in the event expressed by the predicate. Interestingly,

the function-argument opposition is transversal to different approaches to meaning, and it is shared both by model-theoretic and conceptualist semantic theories.

Argument structure is often closely related to “strict” lexical selection (cf. Grimshaw 1990). In this case, semantic arguments are only those that are obligatorily required by the predicate, while adjuncts or circumstantial modifiers are excluded from the lexically projected argument structure. Alternatively, in Fillmore’s *frame semantics* the meaning of a word corresponds to a functional structure consisting of an experienced-based conceptualization of the world, or *frame* (Fillmore and Atkins 1992). The latter is a schematic knowledge-structure (largely influenced by similar notions in cognitive psychology and artificial intelligence, e.g. Minsky’s frames, Shank’s scripts, etc.), which represents a stereotyped situation in terms of its prototypical participants, called *frame elements* (Fillmore *et al.* 2003). For instance, in FrameNet – probably the largest available computational lexicon developed using frame semantics – the core meaning of the verb *buy* is represented by associating this lexical item to a COMMERCIAL_TRANSACTION frame, in turn defined by the following elements: BUYER, GOODS, MONEY and SELLER. Actually, besides these core elements the semantic frame expressed by a lexical item may also contain non-core elements, such as PLACE, TIME, INSTRUMENT, MANNER, PURPOSE, etc.

Although frame elements and thematic roles are cognate notions, since both refer to relations between an event and its participants, they must be distinguished. Roles like BUYER or SELLER can only be interpreted within the specific conceptual space defined by a schematic conceptualization of “buying” situations. A general role like AGENT is instead located at a higher level of abstraction, as a result of generalizing over different event-specific frames. Actually, thematic roles are now recognized as having a highly graded structure, and showing prototypicality effects. Since Dowty (1991), they are typically defined as clusters of more basic semantic properties (e.g. volition, sentience, etc.) that can not be reduced to sets of necessary and sufficient conditions. In FrameNet, frame element labels act as shorthand for common verb properties as emerging from actual verb usages. Frames are then interlinked by different types of relations, so that the overall frame repository actually takes the form of an inheritance network. This intends to capture the fact that situations are conceptualized by lexical items at different degrees of abstraction,

and under various perspectives (cf. for instance *buy* vs. *sell*). Semantic frames thus provide the basis for the representation of a wide array of lexically-related inferences as directly deriving from the stereotyped structure of events and situations expressed by predicates. Moreover, semantic similarity relations are represented by the fact that different lexical items may share the same semantic frame. A significant gap still remains, however, between the unstructured and intuitively chosen labels used in FrameNet and their formal characterization within the structure of interrelated actions and relations forming the semantic frames. The explicit representation of such frame semantic information is a necessary precondition for FrameNet's potential use in text understanding and inference to be fully realized (Chang *et al.* 2002).

In Jackendoff's *Lexical Conceptual Structure* (LCS) predicate argument structures and thematic relations are defined as combinations of more cognitively primitive functional operators. LCS is an example of a *decompositional* approach to lexical ontology, in which the conceptual symbols representing word meanings are decomposed in more elementary and basic elements. Rather than carrying out lexical decomposition in terms of feature lists (e.g. MALE, ANIMATE, ADULT, etc.), LCS assumes a distinction between basic ontological categories (STATE, EVENT, PATH, PLACE, OBJECT, etc.) and conceptual primitive functions (BE, STAY, GO, EXT, ORIENT, CAUSE, etc.), which project ontological categories onto other more complex categories and represent the major members of a family of core functions around which situations (States and Events) are organized (Jackendoff 2002: 363). Conceptual functions provide the building blocks out of which the meaning of lexical items, and in particular their argument structure, can be reconstructed. For instance, the transitive verb *enter* is assigned the LCS [*Event* GO ([*Object* X], [*Path* TO ([*Place* IN ([*Object* Y])])])]), where GO is a primitive conceptual function expressing the movement of an object along a path, and X and Y are typed variables restricting the conceptual categories of the potential argument fillers. Like thematic roles, LCS primitive functions also need an explicit interpretation and precise criteria for their selection. Their status as basic building blocks is however essentially regarded as temporary, and even a basic conceptual symbol like CAUSE is expected to be defined in terms of more primitive representations. LCS symbolic structures on the one hand represent inferential and combinatorial properties of lexical items (e.g. argument

structure), but on the other hand are designed to provide the interface between linguistic structure and other cognitive representations, in particular spatial ones. In fact, the conceptual functions are interpreted onto *Spatial Structures*, where the latter are defined as the encoding of the spatial understanding of the physical world (Jackendoff 2002: 346). This represents an interesting hypothesis about the relationship between lexical ontologies and other possibly non-symbolic cognitive representations, although Jackendoff emphasizes that LCS symbolic representations, and not spatial ones, carry the burden of encoding grammatically relevant properties of language.

The distinction between functions and arguments as a primary aspect of the lexical space plays a key role in formal implementations of the *principle of compositionality*: complex semantic representations are in fact modeled as recursive symbolic structures obtained by means of applying the functions expressed by predicate lexical items to their arguments, according to the syntactic structure of the sentence. As Frege (1923) claims, we could not understand new meaningful expressions if we could not distinguish parts in the thought corresponding to the parts of a sentence. At the same time, the functional structure of semantic representations also raises the issue of the way predicative and argument lexical items interact. In fact, the way meaning composition is represented in most semantic theories takes the form of a strictly *unidirectional* process of functional application. The sharp lexical dichotomy between words acting as semantic functions and words filling argument positions mirrors the opposition between “*active*” concepts (setting number and types of items they can combine with) and non-predicative “*passive*” concepts, filling in argument slots and satisfying type restrictions (Pustejovsky 1995: 39). However clear in its general outline, this model of lexical composition deals only clumsily with polysemy and lexical creativity. Consider, for the sake of concreteness, the word *school*: a school can be built or destroyed (like a house or a car), founded (like a bank or an institution), begin or end (like Summer or a vacation), go on holiday or win a championship (like a group of friends or a football team), be boring or interesting (like a book or a story), etc. Since the predicates that can be applied to *school* are maximally orthogonal with respect of the semantic type of their arguments, explaining the distribution of *school* in terms of the conceptual categories that it expresses would thus entail multiplying the number of its senses. This problem is addressed in Pustejovsky’s

theory of the *Generative Lexicon* (GL) by proposing alternative ways of conceiving both the internal structure of concepts and the way concept composition is modelled (Pustejovsky 1995).

One of the tenets of GL is the rejection of two classical assumptions: that compositionality is a unidirectional functional application process, and that the lexicon is partitioned into a class of active, selection-imposing functional elements and a class of passive, selection-satisfying argument lexemes. GL assumes that even those lexical items that superficially behave as arguments have a complex internal predicative nature governing their linguistic distribution. This generalized predicative structure of lexical items is formally represented as a 4-dimensional *Qualia Structure*. Together with the *Argument Structure*, the *Event Structure* and the *Lexical Inheritance Structure*, the qualia structure provides word entries with a complex multi-layered representation of their content, thereby making “all lexical items as relational to a certain degree” (Pustejovsky 1995: 76). Qualia structures appear in the semantic representation of all the major lexical categories, although their most direct and clear application is in the analysis of nominal concepts, where AGENTIVE and TELIC qualia encode the (proto)typical events in which the referents of nominal concepts are involved.

The qualia structure is intended as a “lexicalization” of a grammatically relevant portion of the contextual knowledge about the entity denoted by a word. According to this view, the fact that a book can be read as part of its typical function, or written as its mode of creation, or similarly the fact that the typical function of a violinist is to play and the one of a knife is to cut are an integral part of the information content lexicalized in the concepts expressed respectively by *book*, *violinist* or *knife*. This in turn means that lexical concepts do not only include properties that can be modelled as monadic features (e.g. shape, color, dimension, etc.), but also information referring to the events and situations in which these entities participate, represented in terms of polyadic predicates. Crucially, in GL the pieces of contextual knowledge encoded in the qualia structure “provide the jumping off point for operations of semantic reconstruction and type change” (Pustejovsky 1995: 77) and are therefore the basis to explain a wide array of creative and dynamic aspects in the lexicon. In fact, systematic lexical polysemy is treated dynamically as the result of an on-line process of lexically-controlled sense creation. Lexical semantic types are fairly *underspecified*, and actual senses are generated

in context as a result of combining and enriching these underspecified semantic structures. It is mainly the information in the qualia structure that drives (and, at the same time, constrains) metonymic reconstructions, sense extensions, adjectival polysemy, etc. These phenomena are modelled through the introduction of more complex modalities of lexical composition, such as *co-composition*, *type-coercion*, *selective binding*, etc. When two lexical items are combined (e.g. a verb and its object or a noun and an adjectival modifier), their qualia structures interact in a complex way and generate context-specific interpretations.

The composition of such semantically articulated lexical items involves merging the information encoded on their various representation layers, and the qualia roles in particular. The notion of qualia structure as a key component in modelling the compositional behaviour of the semantic lexicon raises important issues concerning its proper status and definition. A first salient aspect of qualia is their acting as *structuring dimensions* of word content. In fact, qualia are not themselves meaning components, and under this respect they radically differ from semantic features or primitive conceptual functions. They rather intend to provide a multidimensional structured partition in the space of properties that constitute lexical concepts. The second crucial aspect is that the qualia structure is grounded on the assumption that it is possible to select a portion of contextual knowledge about a category of entities as constitutive of its concept. This point is not uncontroversial and raises the non trivial problem of modelling the qualia structure so as to properly carve out this relevant portion. For instance, the prominence assigned to the TELIC and AGENTIVE roles in the theory implies that these two dimensions are major structuring dimensions in the concept space. Pustejovsky (2001) proposes an ontology of nominal semantic types in which the main partition is between *natural categories* (e.g. dog) and *functional categories* (e.g. knife), with the latter characterized by TELIC and AGENTIVE information, which is instead missing in natural types. This is surely consistent with the bulk of psycholinguistic and cognitive evidence supporting a different representation and organization of artifactual vs natural categories, with functional information playing a key role in identifying the former. In turn this raises the issue of how we characterize the TELIC dimension, since the simple definition as “purpose and function of an entity” is too loose and abstract. Qualia are in fact defined as those aspects of word content that are necessary to capture its lexical polymorphism,

e.g. polysemous alternations, metonymic reconstructions, etc. The point is that a “metaphysical” definition of the qualia may not be perfectly in line with their contribution to the explanation of lexical dynamics. A certain entity may in fact have a proper function in the outer world, without being necessarily the case that this function is relevant to explain its linguistic behavior (cf. Asher and Pustejovsky 2000: 16).

GL addresses the crucial issue of the relationship between the word lexical content and the context, with qualia acting as a sort of *interface*, or a *filter*, between the two. The classical view models this relationship by assigning the context an essentially *selective* role: a word has a fixed number of senses, and then the context *decides* and *selects* an appropriate one. In contrast with this view, the main tenet of GL is that this model is inadequate because it ends up regarding the various senses of a word as all equally distinct, thereby producing their unwarranted multiplication. In GL, the relationship between lexicon and context is somehow reversed. It is an inherent property of the qualia dimensions to act as a structuring filter on contextual knowledge about a category of entities, to single out those aspects that will enter into the concept constitution. However, as Jayez (2001) argues, it is controversial whether this “context-selectivity” of lexical item can be really reduced to the qualia roles - at least as they are currently defined - because they are at the same too loose and too rigid to capture the multidimensionality of concepts.

3. Interlude: word knowledge and word usage

Lexical ontologies rest on the assumption that modelling meaning essentially amounts to representing human lexical competence as fundamentally independent of the way words are used in context. This theoretical stance is reminiscent of the traditional *competence vs. performance* opposition, typical of the generative paradigm in linguistics. In syntactic theory, the opposition enforces an irreducible dichotomy between what we know about the sentences of a language (*i.e.* its grammar) and what we do with them, *i.e.* how we use them in concrete communicative scenarios. In turn, this entails that grammatical descriptions are independent of any use distribution: grammar represents exactly those aspects of language that are supposed to hold true in the mind of an idealized speaker, no matter how language is actually used. In other terms, the representation of such knowledge must be abstracted away from any particular system of use (Townsend and Bever 2001: 37).

Similarly, most symbolic approaches to lexical representations seem to assume a parallel dichotomy between *word knowledge* and *word usage in context*. A lexical ontology of conceptual symbols intends to be a representation of what we know about the information content of a word, irrespective of its real use. Various aspects of word meaning are then explicitly represented in terms of the form and structure of the ontology. The competence vs. performance dichotomy, however, leaves open the issue of how the system of rules and principles defining human grammatical competence is actually used. The same holds of lexical ontologies, which do not lend themselves to modelling lexical dynamics (and, more generally, the way lexical meanings are put to use in real contexts) in a natural way. One major problem concerns the extent to which ontologies are able to account for the conceptual space defined by a word, in the face of the ubiquitous problem of lexical polysemy. In fact, polysemy is typically modelled by encoding it through the system of concepts described by the ontology, with the result of multiplying the repertoire of sense distinctions assigned to a word. As we saw above, GL represents an interesting attempt to explain semantic polymorphism at the level of word interaction in context, rather than as part of abstract lexical representations. According to this hypothesis, the lexical competence also includes generative mechanisms of sense creation. Interestingly, this solution strongly relies on smoothing the boundaries between lexical representations and context by assuming that the qualia structure also encodes important information on prototypical events or situations in which entities appear.

Another example of the limits of ontologies as models of lexical competence concerns their use to express predicate selectional restrictions. The classical hypothesis that one conceptual class should be made to contain all and only the lexical fillers of one predicate argument position is demonstrably too strong in the face of real text evidence. One could nonetheless fall back to the weaker hypothesis that a class of argument fillers be expressed in terms of a disjunction of taxonomy nodes. In fact, even this weaker assumption is often unwarranted. Montemagni and Pirrelli (1998) give examples where predicates selection restrictions are not distributed evenly over taxonomy nodes but rather cut them across, following generalizations that are orthogonal to taxonomical structures. When this is the case, a taxonomy provides virtually no means of generalising over the set of typical arguments of a verb.

It is important to appreciate that this is not a problem of granularity of concept or sense distinctions. As already observed in section 2.2, it is pointless to multiply the senses of *school* to account for its distributional behaviour in terms of strict lexical composition. By the same token, letting finer grained concept sub-hierarchies slip in the semantic type model does not get around the inadequacies of taxonomical information in representing selection restrictions. The problem here seems to be much deeper. Taxonomies need be anchored to fixed classificatory respects, reflecting the requirements of a certain theoretical perspective *P* (Bartsch 1998). It is *P* that allows us to form taxonomies that do not contain cross-classifications. To the contrary, lexical (co)-selection seems to require the ability to pick up relevant respects of similarity *on the fly*, *i.e.* on the basis of dynamically changing context requirements. These two modes of lexical organization, taxonomical and usage-driven, are fundamentally divergent under normal communicative circumstances.

3.1 Semantic types and semantic tokens

These and other open issues in lexical semantics strike us as a direct consequence of regarding the lexicon as a model of word *semantic types*, fundamentally independent of their concrete usages as *tokens* in text contexts. According to this classical view, two different occurrences of an unambiguous English word in context (say *hammer* for example) are regarded as two instances of the same semantic type and thus assumed to be associated with the same invariant meaning core.

In fact, as we shall see in more detail in the following section, some usage-based models of word meaning representation tend to reverse the founding relation between semantic types and semantic tokens (see Elman 2004, for an example of this attitude). Accordingly, the abstract representation of the meaning of a word (*i.e.* its semantic type) is assumed to be based on a constellation of different (but somewhat related) *meaning facets* exhibited by its concrete tokens, rather than *vice versa*. Admittedly, their range of variation is not arbitrary or chaotic, but reflects some contextually-induced variations in a systematic, interpretable way. Novel meaningful extensions in the use of - say - *hammer*, must be coherent with at least *some* other usages of the same word. Nonetheless, it would be impossible to pre-determine the full range of a word's semantic variation *a priori*, as new contexts may provide surprisingly novel constitutive perspectives on word meaning. In this respect, a decontext-

tualized notion of word meaning is nothing but an abstraction, although a fairly convenient one. Accordingly, lexical acquisition does not consist in learning about thousands of separate word meanings, but in constructing a multi-dimensional, densely interconnected semantic space. On-line interpretation of words or phrases is understood as the process of embedding words and phrases in that space. It is to a more careful consideration of the consequences of these usage-based, distributional approaches to word meaning acquisition that we turn now.

4. Bootstrapping meaning from text

Perhaps the most radical departure from symbolic approaches to meaning representation is the Wittgensteinian assumption that lexical knowledge is just a reflection of language usage. Over the last few years, in step with recent advances in understanding language acquisition through computational and robotic models (Broeder and Murre 2000), this assumption has spawned a number of often competing models of machine language learning aimed at investigating the relationship between meaning and concept (see, among others, Luc Steels' contribution in this volume). In the remainder of the present paper we shall mainly be concerned, for reasons that will be clearer in a while, with the interpretation of language usage as a distribution of words in a text. In particular, we intend to investigate in some detail the idea, commonly referred to in the literature as *distributional hypothesis*, that meaning relations are bootstrapped directly from word distributions in large corpora.

4.1 The distributional hypothesis

Since Harris (1968), distributional information about words in context has been taken to play an important role in explaining several aspects of the human language ability. The role of distributional information in developing representations of word meaning is now widely acknowledged in the literature. The distributional hypothesis has been used to explain various aspects of human language processing, such as lexical priming (Lund *et al.* 1995), synonym selection (Landauer and Dumais 1997), retrieval in analogical reasoning (Ramscar and Yarlett 2003) and judgements of semantic similarity (MacDonald and Ramscar 2001). It has also been employed for a wide range of natural language processing tasks, including word sense and syntactic disambiguation, document

classification, identification of translation equivalents, information retrieval, automatic thesaurus/ontology construction and language modelling (see Manning and Schütze 1999 for a comprehensive overview), and has been taken to play a fundamental role in language acquisition (Redington *et al.* 1998).

For our present concerns, the distributional hypothesis makes two basic assumptions. First, the meaning of a word is thought to be based, at least in part, on usage of the word in concrete linguistic contexts. If this is the case, it should then be possible, in principle, to automatically acquire the meaning properties affecting the distributional behaviour of a word, by inspecting a sufficiently large number of its contexts of use. This set of context-sensitive properties provides us with a corpus-based characterisation of the possible meaning facets of a lexical unit. From a more cognitive perspective, an investigation into the process of inducing word meaning properties directly from how words co-occur in context is likely to shed light on the way humans exploit contextual information to learn word uses.

The second underlying assumption concerns classification of word meanings. There is a strong presumption that it is possible to bridge the gap between the wayward information of word usages available in local contiguity and the way people appear to organize this information into bodies of structured knowledge, after being exposed to large amounts of experience. Given any two words *A* and *B*, if they are mutually substitutable in a variety of different contexts, then one can reasonably infer, so the argument goes, that their meanings are similar and ultimately associated with the same semantic class. If true, this hypothesis has the potential of projecting language habits onto linguistic categories, thus establishing a fundamental continuity between ontological knowledge of word meanings and knowledge of how they are used in real contexts.

A qualification is in order at this point. The word *context* will be used here to only mean the narrow linguistic context, i.e. the words uttered/written before and after the word in question. Another possible sense of context, as pragmatic or situated context, including any extra-linguistic information available to the speaker through the particular situation where a sentence is uttered, is not in the focus of the present survey.

This restriction is not totally unreasonable. Firstly, it is certainly true that literate persons acquire many new words through reading, where

nothing but a linguistic context is available (Miller and Charles 1991). Moreover, extra-linguistic evidence, taken alone, may be insufficient for the acquisition of most of our vocabulary: mundane words like *think*, *idea*, *virtue* or even *bachelor* seem to lack stable experiential and observational correlates (Snedeker and Gleitman 2004). Thus, while the text is only of limited help to the acquisition of perceptual correlates of word meanings, it is nonetheless the most natural place to look at when one wants to bootstrap linguistic contingencies for word use such as the meaning components of an abstract or functional word, or the semantic restrictions that words impose on grammatical usage. The distributional hypothesis promises to bootstrap semantic knowledge directly related to this level of linguistic expertise.

5. Algorithmic approaches to meaning bootstrapping

By way of illustration of the distributional hypothesis, we shall briefly be concerned here with a number of different machine learning approaches to meaning acquisition from written texts. The overview is not intended to delve into the technicalities of each such approach, but to mainly emphasize its potential contribution to issues of lexical representation and acquisition. It is important to appreciate, at this stage, that while most of these techniques share many assumptions, they widely differ in the way input words are represented and eventually clustered. One of the most significant merits of the distributional approach to lexical bootstrapping lies, in our view of things, in the variety of perspectives on word representation and classification entertained by the different algorithms. The design, development and assessment of a full-fledged bootstrapping algorithm force developers to make explicit methodological assumptions, and throw in sharp relief issues that are often neglected, if not dismissed, by the symbolic literature on the problem. As we shall see in the remainder of the present section, these assumptions not only point to theoretically interesting aspects of meaning acquisition but also shed novel light on classical problems of word representation. The wealth of these assumptions provides a neat example of the relevance of machine learning research to issues of lexical modelling.

Algorithmically, the distributional hypothesis requires three basic ingredients: i) a computable *representation* of the use of a word in context; ii) the definition of a suitable *distance function* measuring how close any two such representations are; iii) a way to turn distance relations into

similarity-based partitions. In what follows, we shall describe these ingredients in some detail.

5.1 Representing word contexts

Under the distributional approach to word meaning, measuring the semantic similarity between any two words is equivalent to measuring the degree of overlapping between their *sets of linguistic contexts*. But how is the set of contexts of a word to be represented? It is useful to distinguish two different answers to the question. So-called *distributed* representations define the set of contexts of a target word A as a multi-dimensional vector, whose individual components participate in the representation of any other word, but do not precisely point to specific meaning perspectives or features. As we shall see, both distributional vectors and vectors of activity patterns in neural networks are successful examples of distributed representations of this kind (see sections below). On the other hand, *structured representations* try to exploit patterns of the syntax-to-semantics mapping in context, by selectively looking for those co-occurrence patterns only which are likely to be significantly associated with semantic relationships of some kind: for example, a certain verb-noun or adjective-noun construction.

5.1.1 Distributed word representations

Distributional vectors. Most commonly, the behaviour of a target content word A is represented as a high-dimension vector of word co-occurrence frequency distributions. Each vector dimension says how often A is seen in the company of another word in a reference text corpus also called *training corpus*; this figure is possibly divided by A 's overall frequency f_A to normalize differences in token frequency between target words. Thus, the overall number of vector dimensions has an order of magnitude proportional to the size of the vocabulary expressed by the training corpus: vectors with tens or hundreds of thousands of dimensions are fairly common in the literature.

Typically, two words A and B are said to keep company with each other if they appear in the same text span, defined as a sequence of n words. The value of n can vary depending on the approaches, ranging from 1 (corresponding to only immediately adjacent words), to any logical division of the training corpus (a sentence, a paragraph, a document etc.). A

commonly used text width is 10 words, to preserve locality, while minimizing the variability in length of syntactic constructions. In carrying out practical tasks, however, the value n is often empirically adjusted to optimize performance (Li *et al.* 2000). Finally, word strength co-occurrence can be sensitive to the number of intervening words separating A and B , to mimic the correlation span of human working memory (Burgess and Lund 1997): more distant words are perceived as less closely related than more adjacent ones.

A somewhat related, but technically different approach to distributional representations of word usage, known as *Latent Semantic Analysis* (LSA), was originally developed in the context of information retrieval (Landauer and Dumais 1997). Unlike with word co-occurrence matrices, each cell of an LSA distributional matrix counts how many times a given word is found in a particular document. The distributional behaviour of a word is thus defined in terms of a distribution vector of that word over a set of many thousands of training documents (about 30,000). Words which tend to appear in the same documents are considered semantically related.

As an alternative to distributional vectors, A can also be represented as an array of activation values in the units forming the hidden layer of a recurrent Artificial Neural Network (ANN) facing a word prediction task, when A is encoded in the input layer (Elman 1993, 2004). The representation is context-driven and inherently distributed. Different target words correspond to different patterns of activity over the same group of hardware units in the hidden layer. Moreover, these patterns always reflect, together with the current input, the prior state of the network, due to the persistent activation, in input, of the hidden layer activated at the previous time tick. As a result, it is impossible to point to a particular node in the hidden layer where the memory for a particular target word is stored. Furthermore, the representation has the potential to reflect, besides the influence of the target word A on the ensuing context, the history of the short-range dependencies introducing A .

In spite of their commonalities, there is an important respect in which distributional vectors and neural activity patterns are nonetheless quite different representations. When we build the co-occurrence vector for a target word A , we calculate A 's distribution *across all its possible training contexts*. In this respect we are representing the behaviour of A as an average *word type*. Incidentally, this view is shared by most structured

context representations. A node activity array of the word A , on the other hand, is obtained by looking at the activation state of a hidden layer upon seeing a specific *instance* of A embedded in a specific context. This activation, as we saw, inevitably reflects the sequence of words preceding A and will be different from the activity pattern prompted by the same A occurring in a different sentence. This means that a node activity pattern is a representation of a specific occurrence of A in context: that is, a representation of A as a *token*, not as a type.

Structured word representations. In all word representations considered so far, vector dimensions are taken on a par, as basically unstructured defining factors. The extent to which a particular dimension d_j is relevant to A 's representation uniquely depends on the number of times a given word B occurs within the context of A . No assumption is made about the specific relationship between the two words in the text (e.g. whether the two words appear in the same context because they are related through a certain syntactic relation, or because they concern the same topic, etc.). In this specific sense, we can say that representing words on the basis of a distributed representation is a truly exploratory data analysis method, presupposing no known underlying data structure. If a structure is ultimately discovered, this is based on similarity relations between data (or between data dimensions) given a certain distance function.

Using syntactically-related words for bootstrapping lexico-semantic knowledge, on the other hand, takes the converse approach. Similarity is based on a predefined notion of *syntactic structure*. For example, it is known that words entertaining symmetric syntactic relations, such as conjunction or disjunction in context, tend to participate in sets of quasi-synonyms (Widdows and Dorow 2002). It is then reasonable to attempt to bootstrap lexico-semantic clusters by capitalizing on this reliable syntax-to-semantics mapping. More generally, linguistically structured representations tend to take into account those aspects of textual usages, such as word order, syntax and, possibly, rhetorical structure, which are utterly ignored by distributed representations.

Different pieces of linguistic information can be taken into account to characterize syntactic contexts. We can illustrate this point by mentioning the case of information about prepositions. Given a binary syntactic dependency $r \langle w_1, w_2 \rangle$, where r denotes the syntactic relationship hold-

ing between w_1 (the lexical head) and w_2 (its dependent), extracted syntactic contexts can either be encoded in terms of a generic relation holding between a head (be it a noun, a verb or an adjective) and its prepositional complement (Grefenstette 1994) or carry information about the preposition introducing the complement (see Gamallo *et al.* this volume). To be more concrete, from a binary dependency like *in*<*confidence*, *future*> (corresponding to the phrase *the confidence in the future*) the following syntactic contexts of the head noun can be extracted:

1 *confidence*: <P_COMP, *future*>

2 *confidence*: <IN, *future*>

where 1 abstracts away from the preposition type introducing the complement, whereas 2 explicitly encodes this information.

Another important element of variation concerns the type of syntactic relations to be included in the context representation of a word. In particular head-argument pairs (especially verb-noun pairs) have largely been used to represent the meaning of an argument as the set of its governing predicates, or, conversely, to represent the meaning of a predicate as the set of its typical arguments (Lin 1998, Rooth *et al.* 1999, Allegrini *et al.* 2000, Gamallo *et al.* this volume). Another interesting open issue is whether a representation of w should be limited to words governed by w , or if it could also encompass words syntactically governing w . Turning back to the example above, is it only the case that *future* has to be listed among the syntactic contexts of *confidence* or can *confidence* also be listed as a syntactic context of *future*? The co-composition hypothesis (Pustejovsky 1995), which states that in a head-complement dependency also the complement imposes constraints on the head, gives reasons for including, among the semantically-relevant contexts of w , also the words governing w . It goes without saying that choice of a particular type of context representation may have significant repercussions on the typology and quality of acquired semantic information.

5.2 Measuring distances

Given any two real-valued vector representations, we can compute their *Euclidean distance* as the *norm* of their difference, *i.e.* the square root of the sum of quadratic differences of their dimension values as

follows:

$$|\vec{x} - \vec{y}| = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (1)$$

One can then express *vector similarity* as an inverse function of the Euclidean distance, for example: $sim(x, y) = 1/(1+|\vec{x} - \vec{y}|)$. For our present purposes, this reflects the assumption that topological proximity in the space of target words is indicative of semantic closeness.

While real-valued vectors (such as activity arrays of hidden layer units or signal vectors in speech processing) have a straightforward interpretation in terms of spatial relationships, it is less clear how we should interpret them when vector dimensions take as values frequency distributions or probability densities. Other information theoretic metrics are more suitable in this case, such as the so-called *KL divergence* (and other related measures such as *information radius*), saying, for any two distributions q and p , how well q approximates p , or how much information is lost if we use distribution q when the true distribution is p .

It is important to be careful in selecting the right type of probability distribution for the task at hand. Clearly, if we intend to build part-of-speech classes to smooth stochastic language models, the distribution of choice should reflect short-range correlation between sequences of words in a large corpus. Brown *et al.* (1992), for example, use *mutual information* to measure the degree of bi-gram cohesion, and cluster together those words whose merger minimizes reduction in mutual information. On the other hand, long-range correlations, measured within larger text windows, are more likely to give relevant information about the way words semantically co-select one another in context. In general, this explains why different tasks impose different requirements on the width of the text window where we expect words to co-occur.

If we are not entirely confident that our probability distributions tend to converge on reliable estimates of word behaviour, then it is probably wiser to threshold them into Boolean values. Instead of counting how often A is found in the company of B , we simply record that this event either occurs at least once in our corpus (1) or never does (0). For example, following Allegrini *et al.* (2000), we can represent functionally annotated verb-noun pairs as graphs, where each node corresponds to a word type, and a labelled arc between any two nodes represents the type of syntactic relationship linking them at least once in a text corpus. If we call $AV(n_i)$ the set of different arc-verb pairs which n_i combines with in the reference corpus, it is natural to consider the semantic similarity of any two nouns

n_j and n_k as proportional to their *Jaccard coefficient*, defined as follows:

$$\frac{|AV(n_i) \cap AV(n_k)|}{|AV(n_i) \cup AV(n_k)|} \quad (2)$$

Similarly, given a set A of nodes and a set $N(A)$ of their neighbours, that is the set of nodes linked to any $a \in A$, Widdows and Dorow (2002) defines the best new node as the one maximizing the *affinity score* of equation 3 below, a measure highly reminiscent of the so-called *overlap coefficient* (Manning and Schütze 1999).

$$\frac{|N(u) \cap N(A)|}{|N(u)|} \quad (3)$$

5.3 The representation of similarity

Representations of the contexts of use of words in a corpus and a suitable distance function are the two basic ingredients for calculating word similarity. At the heart of the distributional hypothesis lies the strong presumption that one can directly use similarity measures to come up with similarity-based word clusters. This is in line with the common observation that the notion of semantic similarity plays a central role in human concept formation and categorization. However central, similarity is nonetheless a very elusive notion. Determining the conditions under which it can be defined algorithmically can be exceedingly difficult. There is nothing like overall similarity that can be universally measured, but we always have to say in what respect two things are similar. This brings us to the issue of what feature dimensions have to be considered relevant to our categorization judgements. As we shall see in this section, we can identify two main perspectives on similarity-based word classes: a *global perspective* and a *local* one.

Mapping the word semantic landscape from a global perspective means identifying classes according to a fixed set of similarity dimensions, namely those that are found descriptively more salient given an intended domain and some predefined goals. This is the approach taken by any scientific endeavour. It makes eminent sense insofar as we can reasonably keep our perspective anchored to a fixed domain and a set of clearly stated purposes.

Mapping the semantic landscape from a local perspective, on the other hand, means focusing on a particular word w_j , to span the network of

associative relations it entertains with other words. This typically requires continuously updating the similarity perspective given w_j , as the dimensions shared by w_j and w_k can be different from those shared by w_j and another word w_h . Local similarity allows explorative investigation of the semantic neighbourhood of a single word through spreading activation of all admissible associations given a similarity threshold.

In what follows we present three popular ways of representing similarity relations among multi-dimensional vectors of word usage: *clustering*, *topological projection* on two-dimension surfaces and *word graphs*.

5.3.1 Clustering

Clustering techniques assign vector representations exhibiting highest similarity scores to the same cluster. This is done by either incrementally merging similar (*i.e.* close) word vectors into more and more inclusive clusters (*bottom-up or agglomerative clustering*), or splitting an original catch-all cluster into smaller and smaller sub-clusters (*top-down or divisive clustering*). The final outcome of this process is often represented as an upside-down tree (called *dendrogram*), whose root stands for the overall class and the leaves represent individual word types. Branching nodes reflect the history of progressive mergers or splitting: short branches are formed early in bottom-up clustering and late in top-down clustering. Eventually, the obtained word clusters are expected to define (hierarchies of) concept classes.

No matter whether we form clusters bottom-up or top-down, we can do it from either a *global* or a *local* perspective. Let us focus for simplicity on the case of bottom-up clustering. If we take a local perspective, a target word vector A is added to the cluster C if A happens to be the nearest neighbour of one member of C , where the member in question can vary depending on A . This is done through the so-called *single-link* agglomerative clustering. The technique tends to develop elongated clusters, resembling more a chain of related items than a class proper. In lexico-semantic clusters, word chains are the tools of the trade for representing *family resemblances*, whereby clustered items do not seem to exhibit an overall perspective of similarity, but simply form a chain of non-transitive similarity relations, with no common reference prototype, as in the famous Wittgensteinian example of *game*.

Global clustering, on the other hand, tends to maximize within-cluster similarity, by selecting those clusters whose members show the highest

average similarity, that is they are most closely connected to one another. This is equivalent to imagining a cluster as a heap centered around a *cluster prototype member*, which defines a stable set of *core properties* that have to be shared, to different degrees, by all cluster members.

5.3.2 Projection techniques

Another way to represent similarity relations between target words is by means of a topological two-dimension or three-dimension *projection* of high-dimension word vectors. The projection is called topological to mean that it cannot perfectly reproduce the original distances between n -dimensional word vectors (for $n > 3$), but can only approximate them while preserving certain ordering relations.

Methods of dimensional reduction with minimal loss of distance relations among data are generally referred to as *scaling techniques*. They can be used for purposes of data compression and visualization, and include, among others, *factor analysis*, *principal component analysis* and *multidimensional scaling*. Any projection requires, by definition, adopting a *global* perspective from which data are looked at. In practice, this is achieved by finding out some specific dimensions that appear to maximize a global target function. Factor analysis aims at extracting abstract axes (called *factors*) that account for a maximum of correlation between data vectors. Principal components analysis, on the other hand, projects vector representations onto a surface staked out by those axes accounting for the bulk of data variance. This property makes the approach particularly suitable for exploring data spaces defined by poorly correlated vector dimensions. Finally, multidimensional scaling (MDS) works on qualitative data, arranged in an $m \times m$ matrix (where m is the number of input items) directly expressing similarity ranking between input items themselves (elicited on the basis of human judgements). The general idea is to reconstrue the similarity space underlying similarity judgements. This is done by extracting the minimum number of relevant cognitive dimensions able to account for the overall similarity rankings. Clearly, this technique is particularly useful for dealing with psychometric data. More technically, MDS is optimally suitable for treating order invariant data, where metric differences between input data are not so relevant or reliable.

Kohonen's Self-Organizing Maps (SOMs, Kohonen 2001) are also able to visualize topological relations among high-dimension input vec-

tors. A SOM is a usually two-dimension grid of so-called *receptors* (by analogy to specialized neural cells in the brain) that are trained to be increasingly sensitive to input vectors (representing the external stimuli which the grid is exposed to). At the end of a training session, receptors that are sensitive to similar input vectors are located nearby on the map. Unlike scaling techniques, however, SOMs do not aim at reproducing the original distance relations between input data: topologically connected areas of uniformly behaving receptors are often bordered by discontinuous stripes of chaotically behaving receptors. Moreover, as more frequent input similarities call for a larger number of specialized receptors, areas of uniformly behaving receptors are proportional to the frequency with which data similarities are found in the training data. Finally, while projection techniques appear to enforce global constraints concerning the overall distribution of input data, SOMs are trained by *locally* activating and adjusting one receptor at a time, for each input vector. This makes SOMs highly plausible models of neural computation in psycho-linguistically realistic learning conditions.

5.3.3 Word Graphs

Prior knowledge of structure-based regularities can successfully be exploited to build graph representations of semantically-flavoured word relations. Besides relying on structural relations, this family of approaches presupposes existing concept classes, each characterised by one or more prototypical members or *seed words*. Classes are then formed incrementally as nearest neighbour lists of seed words (Widdows and Dorow 2002). Another interesting aspect of these and related techniques is that they don't make use of word frequency distributions (see also Allegrini *et al.* 2000, for a similar approach). This makes this class of algorithms flexible and efficient, but comparatively noise-sensitive. Furthermore, from a cognitive perspective, the main limit of the approach lies in the well-documented sensitivity of speakers memory to word log-frequency effects (Baayen 2001).

6. Comparing the two approaches: knowledge, use and representations

In the previous sections, we have been presenting two radically different ways of representing meanings: *symbolic structures* vs. *patterns of*

usage distributions. It is now time to assess their relative strengths and weaknesses in a comparative way, with a view to addressing at least some of the basic questions we started this chapter with: what type of semantic knowledge can reasonably be bootstrapped from text distributions? To what extent does this usage-based knowledge address the four major *explananda* of any semantic theory, i.e. *meaning multiplicity*, *lexical inferences*, *semantic similarity judgements*, and *meaning composition*? What use can it be put to? And, last but not least, can we expect to bridge the gap between language usage and symbolic knowledge representations?

6.1 What type of knowledge?

Classical, symbolic approaches to word meaning all agree in taking, as their basic units of representation, word semantic types. A word type is supposed to convey the decontextualized, invariant meaning core shared by all its token occurrences in context. Whenever one single meaning core cannot account for the full range of possible systematic usages of a given type, the latter is split into more subtypes, or sense subdivisions, each taking care of a particular partition of usage-based tokens. From this perspective, learning how to use a particular word in context requires i) prior knowledge of the entire battery of its many sense distinctions, ii) locating any such (sub)type within a systematically structured body of explicit semantic relations.

As we saw, the distributional hypothesis reverses the founding relation between semantic types and tokens, by taking the latter as the starting point of a gradual process of *token integration*, culminating in a somewhat distributed representation of types. From this perspective, types are perceived more as collections of usages than decontextualized abstractions of shared meaning cores. These token-derived lexical types are admittedly very different from the traditional notion of lexeme. In fact, distributed vectors tend to conflate grammatical-categorical knowledge (concerning part-of-speech classes such as noun, verb or adjective) with semantic-categorical knowledge (such as the animate/inanimate, or abstract/concrete dichotomy), thus giving a sense of the overall usability of the target word in context, with no specific background information or categorical perspective in mind. This is not necessarily damning. To begin with, distributed representations appear to be fairly robust, as their overall information content gracefully degrades when vector di-

mensions are removed, or text spans are reduced in length. This is an important property in an acquisitional perspective, in the light of the well-known sparseness and incompleteness of word frequency distributions in the child's language input. Moreover, the information conveyed by distributed representations have been shown to be useful for a variety of linguistic tasks, ranging from basic part-of-speech categorization to morphological disambiguation, sense disambiguation and parsing. This lends support to so-called *constraint-satisfaction* approaches to language comprehension (Seidenberg and MacDonald 1999, Tanenhaus and Carlson 1989, among others), allowing more flexibility and communication between language subsystems and calling into question any strict notion of modularity in processing.

Having said that, when it comes to categorical classification, distributed word vectors no doubt provide at best imperfect, suggestive and noisy information. This is not surprising. We already pointed out that word co-occurrence patterns appear to neglect word order, syntax and rhetorical information. All this information has a significant bearing on linguistic categories and knowledge of word meaning. For example, co-occurrence of two adjacent words in context can be considerably less informative if a syntactic constituent boundary intervenes between the two words. Conversely, distant words in a sentence can be tightly related through syntax.

As a remedy to these deficiencies, some scholars have suggested using a syntactically richer notion of context to distil from corpora the semantic information of interest. By looking for the right places in context, we can hope to find only relevant and homogenous semantic information. The last several years of work on the distributional hypothesis have witnessed a prominent shift of focus from a loosely defined notion of context as a simple word sequence to a syntactically structured one. It remains to be seen, however, how helpful syntactically structured contexts are in bootstrapping semantic categories. When we look carefully at distributional evidence in real corpora, there are a number of reasons to be skeptical about this.

First, analysis of word distributions in large corpora lends support to the view that verbs select noun arguments on the basis of varying traits of semantic similarity rather than on membership in homogenous conceptual classes. Typically, lexical fillers of an argument position share one or two such traits, thus exhibiting a sort of family resemblance air. For

example, the word *leaf* is a semantic cognate of *airplane*, *arrow* and *time* relative to *fly* (either in a literal or figurative sense). Likewise, *leaf* makes a selection class of its own with *snow* and *rain* for the verb *fall*. This explains why ontological classes are poor predictors of verb selectional preferences. The latter, unlike the former, appear to define local and contextually-driven perspectives on word meanings and reluctantly fit the global, type-based background enforced by taxonomical knowledge.

Word polysemy makes distributional evidence even more insidious and difficult to interpret. We know that some meaning perspectives are naturally associated with specific sense distinctions and not with others, but we have no way to induce this association from distributed representations of *word types* where all such perspectives are simultaneously taken into account. *School* as a building is defined by a bundle of mainly physical meaning components which are nonetheless poorly correlated with the typical activities where a school as an institution is engaged.

As another potential source of problems, real texts abound with largely prefabricated predicative expressions such as light verb constructions (e.g. *take the shower*), idioms (e.g. *kick the bucket*, *spill the beans*) and other semi-fixed constructions (e.g. *don't get much out of something*, *go a long way*, *I wouldn't put it past somebody to do something*) etc., whose meaning cannot straightforwardly be derived as a function of the individual meanings of their constituent words. Moreover, in common conversational contexts, we do not seem to talk much about events, let alone actions. Rather, we mostly talk about how things are from our perspective (Thompson and Hopper 2001), by making use of a very wide variety of lexico-grammatical constructions such as intransitive-verb, copular and epistemic clauses. Accordingly, very frequent verbs such as *think*, *want*, *say* or *come* put very few limitations on the range of possible contexts where they are used. Surprisingly enough, even in genres such as written narratives where we might expect a higher proportion of clauses with high transitivity, this expectation is not borne out by the data (Hopper and Thompson 1980).

All in all, these observations should warn us away from the risk of trusting corpus evidence blindly, even in the presence of syntactically structured constructions. Language use, as attested through corpus evidence, is a huge haystack with comparatively few semantic needles. Not every needle is good for everything. Some uses and goals require specific approaches to semantic bootstrapping which are not necessarily useful

for other possible tasks. It is important to be extremely clear about what type of knowledge we aim at and for what purposes we intend to acquire it.

6.2 Instructions for use

In what follows we provide the reader with a short list of reasonable correlations between type of knowledge, type of source data and design choices in corpus-based semantic acquisition. The list is largely underspecified, and should only be intended as a very preliminary set of instructions for use, certainly deserving further investigation.

6.2.1 Keeping perspectives under control

Suppose that we are interested in developing a global taxonomy of semantic types in the medical or financial domain. The first thing to say is that such a taxonomy cannot possibly be a “general-purpose” one. Taxonomies presuppose the existence of a fixed perspective, and changing purposes always implies a change of perspective. Even for deceptively simple models such as a taxonomy of house-related concepts, an interior-design standpoint imposes a dramatically different perspective from - say - a house-building approach.

For these reasons, a sensible suggestion is that we carefully select our information sources, in terms of text genres, thematic domains and goal-directed perspectives. Technical or scientific written documents, for example, are more promising candidates than narratives or newspapers articles. Manuals, user guides, blueprints, encyclopaedic texts, domain-specific dictionaries, specialized web pages and the like, tend to convey those word relations only that are consistent with the particular theoretical perspective entertained in the text. In turn, the perspective is bound to minimise word polysemy and sense co-activation, along with irrelevant or inconsistent word associations. Moreover, the conveyed information is expected to cover a particular domain relatively *completely*, with a minimum of construction variability and a maximum of referential transitivity (in the sense of Thompson and Hopper 2001). The fact that the most successful experiments in taxonomy bootstrapping from word distributions we are aware of have been carried out on the basis of technical domains and text genres supports this suggestion.

For much the same reasons, the same text sources can effectively be tapped to acquire the (proto)typical events in which some nominal entities of interest are involved, thus characterising the functional dimensions of such entities relative to a fixed domain or theoretical perspective. Clearly, focusing on some specific domain offers the further bonus of selecting a set of rather technical, information-loaded target words (e.g. *software*, *microprocessor* and *modem*) as opposed to more ordinary words like *milk*, *rain* or *pay*, whose meaning borders are considerably more difficult to carve out due to their frequency of use.

Another practical way to impose static perspectives on text data is to look for words entering specific language patterns only (such as conjunctions, word pairs linked by a specific dependency relation or other meaningful complex constructions like causal expressions), or belonging to a relatively restricted range of semantic categories, as in *named entity recognition*, or domain categories, as in *automated document classification*. The increasing availability of documents annotated with meta-data, finally, offers the further opportunity to look for words in specific marked-up contexts.

All these examples give an indication of how one can selectively use perspectivizing factors to constrain semantic bootstrapping in connection with some practical applications. Unfortunately, to our knowledge, there is very little work on the interaction between design choices in knowledge bootstrapping and issues of lexical knowledge representation, an area certainly deserving more thorough investigation in the near future.

6.2.2 Unleashing perspectives

Suppose now that we are mainly interested in investigating how humans exploit context information to learn, use, categorize and speak about concepts. Different contexts can activate different aspects of word meaning, based on varying dimensions of semantic similarity, that depend, in turn, on the goals and functions that words happen to serve in context. The phenomenon is so pervasive that it soon becomes impossible to provide an effective account of these dimensions independently of the contexts in question.

To give an illustrative example, consider the Italian polysemous word *consiglio* (meaning both ‘council’ and ‘advice’). Table 1 shows the 10 topmost words distributionally similar to *consiglio* considered as a semantic type. Words are ranked by decreasing values of S(ilarity) score

SIMILARITY CHAIN	S SCORE
<i>programma</i> PROGRAM	1.34695e-05
<i>parlamento</i> PARLIAMENT	6.44995e-06
<i>ministero</i> MINISTRY	6.42011e-06
<i>comunicazione</i> COMMUNICATION	5.9057e-06
<i>misura</i> MEASURE	3.45036e-06
<i>riunione</i> MEETING	2.61985e-06
<i>presidente</i> PRESIDENT	2.24574e-06
<i>convinzione</i> CONVICTION	2.14368e-06
<i>assemblea</i> ASSEMBLY	1.91805e-06
<i>persona</i> PERSON	1.71717e-06

Table 1. The topmost similar words of *consiglio* COUNCIL/ADVICE

(in exponential notation), an entropic score based on type frequency distributions (Allegrini *et al.* 2000). This similarity chain highlights the composite nature of relevant word associations and their strong correlation with both senses of *consiglio*. Words marked in bold refer to (different facets of) the council meaning. The other word associations are clearly related to the ‘advice’ meaning. As the two senses are fairly orthogonal, the overall ranking in Table 1 no longer reflects a single similarity gradient, and arbitrarily collapses multidimensional information onto a single axis.

Things dramatically change when distributional evidence is used to identify word associations relative to a specific meaning facet. This information type can be expressed in distributional terms. For instance, the event expressed by a verb can be used to define the background situation

SIMILARITY CHAIN	S SCORE
<i>parlamento</i> PARLIAMENT	0.0019607800
<i>assemblea</i> ASSEMBLY	0.0005764880
<i>riunione</i> MEETING	0.0004528990
<i>persona</i> PERSON	0.0003300620

Table 2. The topmost similar words of *consiglio* relative to the CONVOCARE event

where the entities denoted by the verb arguments perform various functions/roles. We can then consider the different verbs with which *consiglio* occurs in a corpus (possibly with different functional roles) to capture a context-sensitive notion of semantic similarity. In this case, acquired word associations do not refer to *consiglio* considered as a semantic type (as in the previous case), but rather to *consiglio* as a semantic token, i.e. they refer to the specific situation defined by each co-occurring verb (Allegrini *et al.* 2003).

Take for example the case where *consiglio* is the object of *convocare* ('convene'). The list of its topmost similar words is reported in Table 2. Note that all associations identified here are connected with the 'council' meaning, as all associations triggered by other senses and/or meaning facets of *consiglio* are effectively filtered out by the context.

More interestingly, a token-based notion of word similarity is instrumental for exploring the multifarious meaning facets of the 'council' sense of *consiglio*. Tables 3 and 4 illustrate the similarity chains associated with *consiglio* in two different contexts: as the subject of *decidere* ('decide') and as the subject of *deliberare* ('decree'). In spite of the close semantic relatedness of the two verbs, the resulting word associations are remarkably different. Table 3 groups decision-making entities, ranging from financial and governmental institutions to individuals. On the other hand, Table 4 lists assembly-like entities, where decision-making is the result of a collaborative and somewhat institutionalised process. A slight

SIMILARITY CHAIN	S SCORE
<i>Bundesbank</i>	0.0014769100
<i>stato</i> STATE	0.0011207700
<i>governo</i> GOVERNMENT	0.0010024300
<i>operatore</i> OPERATOR	0.0008785130
<i>autorità</i> AUTHORITY	0.0007340040
<i>persona</i> PERSON	0.0006611840
<i>amministratore</i> ADMINISTRATOR	0.0006184290
<i>borsa</i> STOCK EXCHANGE	0.0003935720
<i>organizzazione</i> ORGANIZATION	0.0003113330

Table 3. The topmost similar words of *consiglio* relative to the DECIDERE event

SIMILARITY CHAIN	S SCORE
<i>assemblea</i> ASSEMBLY	0.0060112700
<i>azionista</i> STOCK HOLDER	0.0006839950
<i>governo</i> GOVERNMENT	0.0005555560
<i>operatore</i> OPERATOR	0.0008785130
<i>autorità</i> AUTHORITY	0.0007340040
<i>persona</i> PERSON	0.0006611840
<i>amministratore</i> ADMINISTRATOR	0.0006184290
<i>borsa</i> STOCK EXCHANGE	0.0003935720
<i>organizzazione</i> ORGANIZATION	0.0003113330

Table 4. The topmost similar words of *consiglio* relative to the DELIBERARE event

change in perspective prompts two considerably different lists of semantic associates.

6.3 What can we account for?

We turn back here to what in section 1 we indicated as the major *explananda* of any semantic theory, i.e. *meaning multiplicity*, *lexical inferences*, *semantic similarity judgements*, and *meaning composition*. Our primary goal is to assess the comparative contribution of classical and distributional models of meaning representation to an in-depth understanding of how lexical meaning works. As a subsidiary goal, we also intend to investigate possible perspectives on combining the two approaches.

Meaning multiplicity. As we saw in section 2.1, symbolic approaches easily account for homonymy, but considerably lag behind in representing the polysemous nature of lexical items. The reason of such limitation is twofold, and deeply entrenched in the very nature of this representational paradigm:

- 1 the *discrete* nature of the conceptual symbols used in ontology building does not allow for a straightforward representation of the intrinsic fluidity of lexical meaning in polysemous words. While homonymy is more conducive to being modeled by drawing sharp boundaries between different word senses, each represented as a

distinct conceptual symbol, the same solution is simply not viable to address the systematic and creative usage of polysemous words;

- 2 a deep divide is assumed to separate lexical content and context. Conceptual symbols in lexical ontologies are typically “decontextualized”, and thus face great difficulties in dealing with the “context sensitive” behaviour of polysemous items and other sense creativity processes.

Pustejovsky’s GL model is a most notable exception to this state of affairs, as it places the issues of polysemy and sense creativity at the heart of any satisfactory theory of meaning. Its smoothing the boundaries between lexical content and the context is an attempt to gain some ground towards reaching this goal. Yet, many issues remain open.

Distributional models offer interesting perspectives on meaning multiplicity and polysemy. First, they do not draw discrete subdivisions between word senses. Word representations built from distribution patterns of usage are in fact located in a continuous semantic space. Under this view, polysemy can be regarded as a sort of “emergent property” of lexical items, naturally and spontaneously deriving from the way semantic representations are built as the result of word interactions in contexts. For instance, Kintsch (2001) shows that LSA-style word vector representations can successfully undergo a process of dynamic token adaptation in concrete predicative contexts, thus accounting for polysemous and metaphorical word usages. In general, distributional models can be used as a probe to investigate the dynamics of word meanings, and how polysemy emerges and propagates in the lexicon. In short, while polysemy is a genuine problem for symbolic models, in the distributional approach it is a by-product of the way word semantic representations are acquired.

Lexical inferences. The ability to explicitly represent the inferential potential of lexical items is one of the major strengths of lexical ontologies, which are in fact designed and developed to behave as inferential systems. A key taxonomical relation such as hyperonymy is firmly grounded on lexical entailments. Likewise, semantic role labels can be regarded as a convenient shorthand for representing clusters of lexical entailments concerning the role that an argument takes in the event expressed by a predicate.

In distributional models, the construction of taxonomies from word co-occurrences in texts is a challenging issue, attracting considerable ef-

fort in the NLP and “ontology learning” community (cf. Hearst 1992, Maedche and Staab 2001, Widdows and Dorow 2002; also Faatz in this volume). A common strategy is to take advantage of the particular syntactic patterns where taxonomically-related word meanings typically occur. For instance, harvesting term conjunctions or item lists in texts can be extremely useful to identify co-hyponyms. Results are convincing, when these techniques are applied on technical or scientific texts. On a less positive note, the problem of finding out syntactic patterns that are unambiguously conducive to specific semantic relations has no general solution and is mostly tackled heuristically and through brute force strategies. Likewise, it is far from clear how it is possible to extend the approach to semantic relations other than hyperonymy/hyponymy.

In our opinion, however, distributional models offer another and perhaps even more interesting contribution to the study of lexical inferences. However weak in discovering entailment relations between lexical items, these methods are nevertheless very useful to investigate other types of inferences, especially those based on *analogy*. This is a powerful and central phenomenon in cognition, and a core ingredient in categorization processes and language acquisition (cf. Gentner and Marckman 1995, Tomasello 2000). Analogy is also known to play an active role in metaphorical and sense extension processes (cf. Lakoff 1987). Although they do not find a place in symbolic ontological systems, analogy-based inferences have surely a key role in meaning dynamics. Analogy resides in the ability to individuate structural invariants in the inputs, and actually distributional models look for word distributional invariants in texts. We can thus legitimately expect these methods to be extremely useful to explore these key aspects of semantic inferential competence (Kintsch 2001).

Semantic similarity judgements. In symbolic models, two words are semantically similar to the extent that their independent representations exhibit topological or structural commonalities: e.g. if there is a short path connecting their location in a semantic network like WordNet, or if they share a certain conceptual symbol at a given level of semantic decomposition. On the other hand, in distributional models semantic similarity is anchored directly to word usage distributions and represents a defining principle of word meaning representations: a word meaning is understood only as a place holder in a high-dimensional similarity

space and can thus belong to each and every class to a different degree, depending on its similarity to other meanings. Accordingly, it is not possible to change the position of any individual meaning in the space without changing the full set of relations entertained by all other meanings. This contrasts with lexical ontologies, where semantic similarity can be regarded as a sort of “second order property”, read off a formal architecture that is defined independently of it.

To better understand the differences between these two views, it is useful to compare alternative models of categorization. In the so-called “classical” view, concepts are defined in terms of a set of features, expressing necessary and sufficient conditions for an entity to be an instance of the concept. Exemplar-based models on the other hand conceive a concept as the results of abstracting commonalities out of a set of similar exemplars. Thus, similarity directly enters into the definition of a concept: an entity in an instance of a concept, exactly because it is highly similar to its (most prototypical) exemplars. The distributional approach is strongly related to this exemplar-based view of conceptual representations. A given semantic type is the result of “clustering” together various distributions of its tokens, each representing an exemplar of how the word is used. So, inter-exemplar similarity is the key to the emergence of word meaning from usage distributions. Actually, distributionally-based semantic similarity has proven to be instrumental in assessing the semantic closeness between words for a number of NLP applications (cf. for instance Lee 1997, Curran 2003, among the many others). Besides, its investigation allows us to better understand exemplar-based processes of categorization and the formation of abstract representations, as well as the way they contribute to the the shape of lexical meaning.

Semantic composition. In the symbolic paradigm, the compositional properties of lexical items are explicitly represented by augmenting ontologies with functional symbols for predicative words (cf. section 2.2) Although this type of representation is clearly missing in distributional models, it can nonetheless be recovered indirectly. To give but one example, in the distributional system CLASS (Allegrini *et al.* 2000, 2003), nouns are represented in terms of their distributions as subjects or direct objects of different verbs. This allows the system to develop representations that are sensitive to the typical events nouns participate in. Conversely, verbs are clustered in terms of nouns occurring as their sub-

ject or direct objects. Two verbs are semantically similar to the extent that they occur with similar argument-filling nouns. Conversely, two argument-filling nouns are semantically similar to the extent that they occur with similar verbs. CLASS thus performs a sort of distributional “approximation” of the function-argument opposition. This information is implicitly represented in the resulting word clusters, rather than being explicitly encoded and labelled through conceptual symbols.

In the symbolic paradigm, lack of explicit representation is equivalent to no information at all. There is no sense in which a symbol can function only partially, or be not one hundred per cent explicit or provide only partial information. A symbol must either be there or not. To the contrary, partly explicit, continuously valued representations disclose new attractive perspectives on semantic investigation. As observed in section 2.2, there is widespread consensus that notions like AGENT or THEME are not definable in terms of a discrete set of necessary and sufficient conditions. This view is also shared by cognitive scientists, who view semantic roles as the result of a process of schematization from concrete verb tokens. From various tokens of a certain event, we can abstract some common properties shared by the entities involved in the event. Ultimately, these properties coalesce in schematic abstractions corresponding to the ones linguists use to label thematic roles. As FrameNet shows conclusively, between event specific roles and highly schematic abstractions like AGENT, THEME or PATIENT lies a whole variety of graded schematic roles, mostly eluding classical symbolic accounts. Distributional models are useful probes into the combinatorial continuum of lexical items.

7. Is there a gap to be bridged? Concluding remarks

In this paper, we tried to shed light on issues of representation, acquisition and categorization of lexical meaning from both a theoretical and a computational vantage point. We started overviewing aspects of lexical representations in the classical symbolic paradigm. In particular we observed that word meaning co-selections are dealt with only partially even by the most advanced models of lexical representation. The main reason for this unsatisfactory state of affairs is arguably their lack of flexibility in accounting for the constructive role of context in shaping up word meanings.

The distributional hypothesis is an attempt to tackle this dilemma from the other end. By representing words as distribution vectors, we can hope to discover semantic word classes as clouds of distributionally close points in an n -dimension space. The second related hope is that word classes of this kind can approximate concept nodes in a taxonomy. If supported by empirical evidence, this hope promises to bridge the gap between the goal-directed need for effective and compact systems of semantic knowledge representation (*e.g.* taxonomies) and what we presently know about human strategies for meaning acquisition and use.

There is more than one reason to suspect that, taken at face value, the distributional hypothesis stops short of bridging this gap. Somewhat ironically, this negative conclusion should *not* be intended to imply that the distributional hypothesis is wrong. To understand why it is so, we need to go over some of the technical points we raised in the previous pages.

First, we discussed the assumption that conceptual nodes can correlate with word co-selection in context, to conclude that this is unwarranted. Although it is generally true that quasi-synonyms are mutually interchangeable in context, the converse statement, *i.e.* that mutually interchangeable words are quasi-synonymic, is easily falsifiable. Distributional clusters are apt to capture the elusive notion of family resemblance, with cluster members sharing a couple of interesting semantic dimensions (or traits), possibly based on the selection properties of their predicates. However, word families of this kind are a far cry from coherent ontological classes. The same holds for word clusters based on co-distribution in “word bags”, whose semantic glue is even more unpredictable and heterogenous.

A related but often neglected observation is that a taxonomy requires the preliminary definition of a coherent theoretical perspective or background theory (Bartsch 1998). It is moot that most current usages of the word space metaphor can meet this requirement in a principled way. Unsupervised approaches to distribution-based clustering take all defining dimensions on a par, and try to discover an underlying data structure by maximizing (or minimizing) a global distribution function (for example, the probability that the training data are stochastically generated by the underlying model). Surely, this is a reasonable thing to do under the assumption that the distribution sources (the training texts) i) entertain the intended overall perspective and ii) do not contain any other, possibly

orthogonal, perspective. This is tantamount to saying that distributional evidence is a valuable basis for bootstrapping taxonomical perspectives only insofar as the source texts are *already* generated according to the intended perspective. Narrative texts and ordinary conversations do not normally entertain a fixed theory-driven conceptual perspective. As a result, no underlying semantic structure can miraculously be generated out of them. In this connection, we suggested that a better control of similarity perspectives can be attained by imposing *external* constraints over text domain, topic and genre. The suggestion appears to be supported by recent work on the automated acquisition of domain-specific taxonomies.

A further related observation is that people happen to use highly structured bodies of semantic knowledge such as domain ontologies only *occasionally*, whenever they feel the need for consciously taking a global perspective on their concepts. Under normal circumstances, however, taxonomical relations are simply *not* the primary, let alone exclusive, form of meaning organization in the speaker's conceptual space. Other types of concept associations appear to correlate with human predispositions, abilities, contingencies, occasional needs, interests and preferences. Some of them play an active role in language use to determine word co-selection, disambiguation strategies, meaning extensions, lexical inferences and the like. Some others are certainly rooted in the innate workings of our sense organs and the brain, and are likely to play a major role in child early concept formation, proto-typicality effects in semantic categorization and common sense reasoning.

In any case, the dynamic interaction of such multiple perspectives on concept classification is a real challenge for any theory-laden, type-based approach to word meaning representation. A *sole* is at the same time a fish and a type of food and can accordingly be classified as both a *living* and *not living* entity. Surely we can adjust our sense subdivisions so as to have *sole*₁ (as fish) and *sole*₂ (as food) belonging to distinct nodes in the same taxonomy. However successful this solution may be, the hypothesis that sense subdivisions can be established once and for all stumbles upon a number of both principled and practical difficulties. As we repeatedly pointed out in the previous pages, meaning is inextricably shaped by context. Such a constructive view is seriously under-determined by symbolic lexical representations, as they all rest on the assumption that lexical meaning relations can, to a large extent, be

listed out of context for them to be picked up appropriately when context requires. Lexical creativity calls for more dynamic and graded word meaning representations.

Getting acquainted with the inherent limitations of both symbolic and distributional approaches is probably the best way to make the most of the two paradigms. Of late, the bulk of work on lexical knowledge bootstrapping has focused on the problem of choosing the right sort of *clustering algorithm* or the right type of *distance measure*. This has certainly advanced our understanding of the issue. Yet, it appears that methodological and experimental design issues deserve considerably more attention than they received so far. Furthermore, it is also extremely important to be clear about the sort of output representation we aim at. In the paper, we distinguished the following four different types of perspectives on word categorization:

- a) global & type-based (*e.g.* a classical taxonomy like WordNet);
- b) local & type-based: (*e.g.* a thesaurus);
- c) local & token-based: (*e.g.* context-sensitive similarity chain);
- d) global & token-based: (*e.g.* context-sensitive projection of an n -dimensional word space).

Each of these four types responds to specific goals and requirements and can be put to use for different applications. Classical taxonomies are probably the tool of the trade for building domain models and making adequate inferences about entities in the domain. A thesaurus is more suitable for expanding term-based queries or, more generally, for investigating the semantic neighborhood of a specific concept type. Token-based similarity chains provide, in turn, context-oriented knowledge that proves to be instrumental for carrying out classical NLP tasks or, more generally, context-sensitive reasoning. Finally, “on the fly” word space projections can considerably improve our understanding of the inherent short-scale dynamics of concept learning.

There is one important point that this tentative classification brings home. It makes comparatively little sense to cast local, token-based classifications into global, type-based knowledge structures. Even though this were an attainable goal (but we gave reasons to doubt this), the end result would simply compound the weaknesses of the two approaches,

by casting categorically noisy information into a rigid, decontextualized mould. Likewise, existing ontologies can tell us very little about the fluidity of word meaning, similarity judgements and “on the fly” inferences in context. When things are looked at from this angle, *there is no gap to be bridged*, simply because the two approaches are complementary and respond to radically divergent objectives.

What we seem to be in need of is not a technique to convert one type of representation into another, but rather an integrated theoretical and computational framework accounting for the way speakers acquire, use and extend word meanings, or make inferences and similarity judgements about them. This investigation will have an impact on the forms of meaning representation as we understand them now, and eventually challenge traditional views on meaning computation as a formal manipulation of complex symbols. In the end, we can expect such an integrated, multi-disciplinary framework to spawn usage-based models of meaning representation that squarely address word polysemy and sense creativity as central explananda of semantic theory. From this standpoint, purely distributional approaches to word meaning will have to be used on increasingly richer representations, where surface linguistic structures are paired with, possibly extra-linguistic, meaning correlates. The growing availability of reliable repertoires of word-based semantic information (including terminology, phraseology, named entities, collocates, multimedia perceptual correlates etc.), along with recent advances in robotics, machine learning and robust parsing architectures, all contribute towards making this perspective an ambitious but attainable research goal.¹

References

- Allegriani, P., Montemagni, S., Pirrelli, V. (2000). “Learning word clusters from data types”, in *Proceedings of COLING 2000*, Saarbrücken, Germany.
- Allegriani, P., Montemagni, S., Pirrelli, V. (2003). “Example-based automatic induction of semantic classes through entropic scores”, *Linguistica computazionale*, 16-17: 1-45.
- Asher, N. and Pustejovsky, J. (2000). “The metaphysics of words in context”, ms. Brandeis University.

¹This chapter was jointly developed by the three authors. For academic purposes, Alessandro Lenci bears responsibility for sections 1 through 3, Vito Pirrelli for sections 4 through 6.1 and Simonetta Montemagni for sections 6.2 through 7.

- Baayen, R.H. (2001). *Word Frequency Distributions*, Dordrecht, Kluwer Academic Publishers.
- Barsalou, L.W. (1982). "Context-independent and context-dependent information in concepts", *Memory and Cognition*, 10: 82–93.
- Bartsch, R. (1998). *Dynamic Conceptual Semantics: A Logico-Philosophical Investigation into Concept Formation and Understanding*, Stanford, CA, CSLI.
- Broeder, P. and Murre, J. (2000). *Cognitive Models of Language Acquisition*, Cambridge, Cambridge University Press.
- Brown, P.F., Pietra, V.J.D., DeSouza, P.V., Lai, J.C., Mercer, R.L. (1992). "Class-based n-gram models of natural language", *Computational Linguistics*, 18(4): 467–479.
- Burgess, C., and Lund, K. (1997). "Modeling parsing constraints with high-dimensional context space", *Language and Cognitive Processes*, 12: 177–210.
- Busa, F., Calzolari, N., Lenci, A., and Pustejovsky J. (2001). "Building a semantic lexicon: structuring and generating concepts", in Bunt H., Muskens R., and Thijsse E. (eds.), *Computing Meaning Vol. II*, Dordrecht, Kluwer: 29–51.
- Chang, N., Narayanan, S. and Petrucci, M.R.L. (2002). "Putting frames in perspective", in *Proceedings of the Nineteenth International Conference on Computational Linguistics*, Taipei, Taiwan.
- Curran, J.R. (2003). *From Distributional to Semantic Similarity*, Ph.D. thesis, University of Edinburgh.
- Dowty, D.R. (1991). "Thematic proto-roles and argument selection", *Language*, 67(3): 547–619.
- Elman, J. (1993). "Learning and development in neural networks: The importance of starting small", *Cognition*, 48: 71–99.
- Elman, J. (2004). "An alternative view of the mental lexicon", *Trends in Cognitive Sciences*, 8(7): 301–306.
- Fellbaum, C. (ed.) (1998). *WordNet. An Electronic Lexical Database*, Cambridge, MA, MIT Press.
- Fillmore, C.J. and Atkins, B.T.S. (1992). "Towards a frame-based organization of the lexicon: The semantics of RISK and its neighbors", in Lehrer, A and Kittay, E. (eds.), *Frames, Fields, and Contrast: New Essays in Semantics and Lexical Organization*, Hillsdale, Lawrence Erlbaum Associates: 75–102.
- Fillmore, C.J., Johnson, C.R. and Petrucci, M.R.L. (2003). "Background to Framenet", *International Journal of Lexicography*, 16(3): 235–250.
- Frege, G. (1923). "Logische Untersuchungen. Dritter Teil: Gedankenfüge", *Beiträge zur Philosophie des Deutschen Idealismus vol. III*, 36–51 (Translated as "Compound Thoughts, Logical Investigations", Blackwell, Oxford, 1977: 55–78).
- Gentner, D. and Markman, A. (1995). "Similarity is like analogy: Structural alignment in comparison", in Cacciari, C. (ed.), *Similarity in Language, Thought and Perception*, Brussels, BREPOLs.
- Goodman, N. (1972). "Seven strictures on similarity", in Goodman, N. (ed.), *Problems and Projects*, New York, Bobbs-Merrill: 437–447.
- Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*, Boston, Kluwer Academic Publishers.
- Grimshaw, J. (1990). *Argument Structure*, Cambridge, MA, MIT Press.

- Gruber, T. R. (1993). "A translation approach to portable ontologies", *Knowledge Acquisition*, 5(2):199–220.
- Guarino, N. (1998). "Some ontological principles for designing upper level lexical resources", in *Proceedings of LREC 1998*, Granada, Spain: 527–534.
- Harris, Z.S. (1968). *Mathematical Structures of Language*, New York, Wiley.
- Hearst, M. (1992). "Automatic acquisition of hyponyms from large text corpora", in *Proceedings of the Fourteenth International Conference on Computational Linguistics*, Nantes, France, July 1992.
- Hopper, P.J. and Thompson, S.A. (1980). "Transitivity in grammar and discourse", *Language*, 56:251–299.
- Jackendoff, R. (2002). *Foundations of Language. Brain, Meaning, Grammar, Evolution*, New York, Oxford University Press.
- Jayez, J. (2001). "Underspecification, context selection, and generativity", in Bouillon, P. and Busa F. (eds.), *The Language of Word Meaning*, Cambridge, Cambridge University Press: 124–148.
- Kintsch, W. (2001). "Predication", *Cognitive Science*, 25: 173–202.
- Kohonen, T. (2001). *Self-Organizing Maps*, Berlin Heidelberg, Springer-Verlag, Third Edition.
- Lakoff, G. (1987). *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*, Chicago, University of Chicago Press.
- Landauer, T. K. and 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): 211–240.
- Lee, L. (1997). *Similarity-Based Approaches to Natural Language Processing*, Ph.D. thesis, Harvard University.
- Lenci, A., Bel, N., Busa, F., Calzolari, N., Gola, E., Monachini, M., Ogonowsky, A., Peters, I., Peters, W., Ruimy, N., Villegas, M., Zampolli, A. (2000). "SIMPLE: a general framework for the development of multilingual lexicons", *International Journal of Lexicography*, 13(4): 249–263.
- Levin, B. (1993). *English Verb Classes And Alternations: A Preliminary Investigation*, Chicago, The University of Chicago Press.
- Li, P., Burgess, C., and Lund, K. (2000). "The acquisition of word meaning through global lexical co-occurrences", in Clark, E.V. (ed.), *Proceedings of the Thirtieth Stanford Child Language Research Forum*, Stanford, CA, CSLI: 167–178.
- Lin D. (1998). "Automatic retrieval and clustering of similar words", in *Proceedings of COLING-ACL '98*, Montreal, Canada, August 1998.
- Lin, E.L., and Murphy, G.L. (2001). "Thematic relations in adults' concepts", *Journal of Experimental Psychology: General*, 130: 3–28.
- Lund, K., Burgess, C., Atchley, R.A. (1995). "Semantic and associative priming in high-dimensional semantic space", in *Proceedings of the Cognitive Science Society*, Hillsdale, N.J., Erlbaum Publishers: 660–665.
- MacDonald, S. and Ramsar, M.J.A. (2001). "Testing the distributional hypothesis: the influence of context on judgements of semantic similarity", in *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, Edinburgh, Scotland.

- Maedche, A. and Staab, S. (2001). "Ontology Learning for the Semantic Web", in *IEEE Intelligent Systems*, 16(2).
- Manning, C.D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*, Cambridge, MA, MIT Press.
- Marconi, D.(1997). *Lexical Competence*, Cambridge, MA, MIT Press.
- Medin, D.L., Goldstone, R.L., Gentner, D. (1993). "Respects for similarity", *Psychological Review*, 100: 254–278.
- Miller, G.A. (1998). "Nouns in WordNet", in Fellbaum (1998): 23–46.
- Miller, G.A., Charles, W.G. (1991). "Contextual Correlates of Semantic Similarity", in *Language and Cognitive Processes*, 6(1): 1–28.
- Montemagni, S. and Pirrelli, V. (1998). "Augmenting WordNet-like lexical resources with distributional evidence. An application oriented perspective", in *Proceedings of the COLING-ACL '98 Workshop on Usage of WordNet in Natural Language Processing Systems*, Montreal, Canada, August 1998.
- Pinker, S. (1989). *Learnability and Cognition: The Acquisition of Argument Structure*, Cambridge, MA, MIT Press.
- Pustejovsky, J. (1995). *The Generative Lexicon*, Cambridge, MA, MIT Press.
- Pustejovsky, J. (2001). "Type construction and the logic of concepts", in Bouillon, P. and Busa F. (eds.), *The Language of Word Meaning*, Cambridge, Cambridge University Press: 91–123.
- Ramscar, M. and Yarlett, D. (2003). "Semantic grounding in models of analogy: an environmental approach", *Cognitive Science*, 27(1): 41–71.
- Rooth, M., Riezler, S., Prescher, D., Carroll, G., Beil, F. (1999). "Inducing a semantically annotated lexicon via EM-based clustering", in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, Maryland, USA: 104–111.
- Redington, M., Chater, N., Finch, S. (1998). "Distributional information: A powerful cue for acquiring syntactic categories", *Cognitive Science*, 22: 425–469.
- Saint-Dizier, P. and Viegas, E. (1995). "An introduction to lexical semantics: a linguistic and psycholinguistic Perspective", in Saint-Dizier, P. and Viegas E. (eds.), *Computational Lexical Semantics*, New York, Cambridge University Press.
- Seidenberg, M.S. and MacDonald, M.C. (1999). "A probabilistic constraints approach to language acquisition and processing", *Cognitive Science*, 23: 569–588.
- Snedeker, J. and Gleitman, L. (2004). "Why it is hard to label our concepts", in Hall, D.G. and Waxman, S.R. (eds.), *Weaving a Lexicon*, Cambridge, MA, MIT Press: 257–293.
- Tanenhaus, M.K., and Carlson, G.N. (1989). "Lexical structure and language comprehension", in Marslen-Wilson, W.D. (ed.), *Lexical Representation and Process*, Cambridge, MA, MIT Press.
- Thompson, S.A. and Hopper, P.J. (2001). "Transitivity, clause structure, and argument structure: evidence from conversation", in Bybee, J.L. and Hopper, P.J. (eds.), *Frequency and the Emergence of Linguistic Structure*, Amsterdam, Benjamins: 27–60.
- Tomasello, M. (2000). "Do young children have adult syntactic competence?", *Cognition*, 74: 209–253.
- Townsend, D.J. and Bever, T.G. (2001). *Sentence Comprehension: The Integration of Habits and Rules*, Cambridge, MA, MIT Press.

- Vossen, P. (1998). "Introduction to EuroWordNet", in N. Ide, D. Greenstein, P. Vossen (eds.), *Special Issue on EuroWordNet. Computers and the Humanities*, Volume 32, Nos. 2-3: 73-89.
- Vossen P. (2003), "Ontologies", in Mitkov, R. (ed.), *Handbook Of Computational Linguistics*, Oxford, Oxford University Press.
- Widdows, D. and Dorow, B. (2002). "A graph model for unsupervised lexical acquisition", in *19th International Conference on Computational Linguistics*, Taipei: 1093-1099.