ISA meets Lara: An incremental word space model for cognitively plausible simulations of semantic learning

Marco Baroni
CIMeC (University of Trento)
C.so Bettini 31
38068 Rovereto, Italy
marco.baroni@unitn.it

Alessandro Lenci
Department of Linguistics
University of Pisa
Via Santa Maria 36
56126 Pisa, Italy
alessandro.lenci@ilc.cnr.it

Luca Onnis
Department of Psychology
Cornell University
Ithaca, NY 14853
lo35@cornell.edu

Abstract

We introduce *Incremental Semantic Analysis*, a fully incremental word space model, and we test it on longitudinal child-directed speech data. On this task, ISA outperforms the related Random Indexing algorithm, as well as a SVD-based technique. In addition, the model has interesting properties that might also be characteristic of the semantic space of children.

1 Introduction

Word space models induce a semantic space from raw textual input by keeping track of patterns of co-occurrence of words with other words through a vectorial representation. Proponents of word space models such as HAL (Burgess and Lund, 1997) and LSA (Landauer and Dumais, 1997) have argued that such models can capture a variety of facts about human semantic learning, processing, and representation. As such, word space methods are not only increasingly useful as engineering applications, but they are also potentially promising for modeling cognitive processes of lexical semantics.

However, to the extent that current word space models are largely non-incremental, they can hardly accommodate how young children develop a semantic space by moving from virtually no knowledge of the language to reach an adult-like state. The family of models based on singular value decomposition (SVD) and similar dimensionality reduction techniques (e.g., LSA) first construct a full co-occurrence matrix based on statistics extracted from the whole input corpus, and then build a model at once via matrix algebra operations. Admittedly, this is hardly a plausible simulation of how children learn word meanings incrementally by being exposed to short sentences containing a relatively small number of different words. The lack of incrementality of several models appears conspicuous especially given their explicit claim to solve old theoretical issues about the acquisition of language (e.g., (Landauer and Dumais, 1997)). Other extant models display some degree if incrementality. For instance, HAL and Random Indexing (Karlgren and Sahlgren, 2001) can generate well-formed vector representations at intermediate stages of learning. However, they lack incrementality when they make use of stop word lists or weigthing techniques that are based on whole corpus statistics. For instance, consistently with the HAL approach, Li et al. (2000) first build a word co-occurrence matrix, and then compute the variance of each column to reduce the vector dimensions by discarding those with the least contextual diversity.

Farkas and Li (2000) and Li et al. (2004) propose an incremental version of HAL by using a a recurrent neural network trained with Hebbian learning. The networks incrementally build distributional vectors that are then used to induce word semantic clusters with a Self-Organizing Map.Farkas and Li (2000) does not contain any evaluation of the structure of the semantic categories emerged in the SOM. A more precise evaluation is instead performed by Li et al. (2004), revealing the model's ability to simulate interesting aspects of early vocabulary dynamics. However, this is achieved by using hybrid word

Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition, pages 49–56,
Prague, Czech Republic, June 2007

©2007 Association for Computational Linguistics

representations, in which the distributional vectors are enriched with semantic features derived from WordNet.

Borovsky and Elman (2006) also model word learning in a fairly incremental fashion, by using the hidden layer vectors of a Simple Recurrent Network as word representations. The network is probed at different training epochs and its internal representations are evaluated against a gold standard ontology of semantic categories to monitor the progress in word learning. Borovsky and Elman (2006)'s claim that their model simulates relevant aspects of child word learning should probably be moderated by the fact that they used a simplified set of artificial sentences as training corpus. From their simulations it is thus difficult to evaluate whether the model would scale up to large naturalistic samples of language.

In this paper, we introduce *Incremental Semantic Indexing* (ISA), a model that strives to be more developmentally plausible by achieving full incrementality. We test the model and some of its less incremental rivals on Lara, a longitudinal corpus of child-directed speech based on samples of child-adult linguistic interactions collected regularly from 1 to 3 years of age of a single English child. ISA achieves the best performance on these data, and it learns a semantic space that has interesting properties for our understanding of how children learn and structure word meaning. Thus, the desirability of incrementality increases as the model promises to capture specific developmental trajectories in semantic learning.

The plan of the paper is as follows. First, we introduce ISA together with its main predecessor, Random Indexing. Then, we present the learning experiments in which several versions of ISA and other models are trained to induce and organize lexical semantic information from child-directed speech transcripts. Lastly, we discuss further work in developmental computational modeling using word space models.

2 Models

2.1 Random Indexing

Since the model we are proposing can be seen as a fully incremental variation on *Random Indexing* (RI), we start by introducing the basic features of RI (Karlgren and Sahlgren, 2001). Initially, each context word is assigned an arbitrary vector representation of fixed dimensionality d made of a small number of randomly distributed +1 and -1, with all other dimensions assigned a 0 value (d is typically much smaller than the dimensionality of the full co-occurrence matrix). This vector representation is called *signature*. The context-dependent representation for a given target word is then obtained by adding the signatures of the words it co-occurs with to its *history* vector. Multiplying the history by a small constant called *impact* typically improves RI performance. Thus, at each encounter of target word t with a context word c, the history of t is updated as follows:

$$\mathbf{h}_t \mathrel{+}= i \times \mathbf{s}_c \tag{1}$$

where i is the impact constant, $\mathbf{h}_t$ is the history vector of t and $\mathbf{s}_c$ is the signature vector of c. In this way, the history of a word keeps track of the contexts in which it occurred. Similarity among words is then measured by comparing their history vectors, e.g., measuring their cosine.

RI is an extremely efficient technique, since it directly builds and updates a matrix of reduced dimensionality (typically, a few thousands elements), instead of constructing a full high-dimensional co-occurrence matrix and then reducing it through SVD or similar procedures. The model is incremental to the extent that at each stage of corpus processing the vector representations are well-formed and could be used to compute similarity among words. However, RI misses the "second order" effects that are claimed to account, at least in part, for the effectiveness of SVD-based techniques (Manning and Schütze, 1999, 15.4). Thus, for example, since different random signatures are assigned to the words *cat*, *dog* and *train*, the model does not capture the fact that the first two words, but not the third, should count as similar contexts. Moreover, RI is not fully incremental in several respects. First, on each encounter of two words, the same fixed random signature of one of them is added to the history of the other, i.e., the way in which a word affects another does not evolve with the changes in the model's knowledge about the words. Second, RI makes use of filtering and weighting procedures that rely on global statistics, i.e., statistics based on whole corpus counts. These procedures include: a) treating the most frequent words as stop words; b) cutting off the lowest frequency words as potential contexts; and c) using mutual information or entropy measures to weight the effect of a word on the other). In addition, although procedures b) and c) may have some psychological grounding, procedure a) would implausibly entail that to build semantic representations the child actively filters out high frequency words as noise from her linguistic experience. Thus, as it stands RI has some noticeable limitations as a developmental model.

2.2 Incremental Semantic Analysis

Incremental Semantic Analysis (ISA) differs from RI in two main respects. First and most importantly, when a word encounters another word, the history vector of the former is updated with a weighted sum of the signature *and* the history of the latter. This corresponds to the idea that a target word is affected not only by its context words, but also by the semantic information encoded by that their distributional histories. In this way, ISA can capture SVD-like second order effects: *cat* and *dog* might work like similar contexts because they are likely to have similar histories. More generally, this idea relies on two intuitively plausible assumptions about contextual effects in word learning, i.e., that the information carried by a context word will change as our knowledge about the word increases, and that knowing about the history of co-occurrence of a context word is an important part of the information being contributed by the word to the targets it affects.

Second, ISA does not rely on global statistics for filtering and weighting purposes. Instead, it uses a weighting scheme that changes as a function of the frequency of the context word at each update. This makes the model fully incremental and (together with the previous innovation) sensitive not only to the overall frequency of words in the corpus, but to the order in which they appear.

More explicitly, at each encounter of a target word t with a context word c, the history vector of t is updated as follows:

$$\mathbf{h}_t \mathrel{+}= i \times (m_c \mathbf{h}_c + (1 - m_c)\mathbf{s}_c) \tag{2}$$

The constant i is the impact rate, as in the RI formula (1) above. The value m_c determines how much the history of a word will influence the history of another word. The intuition here is that frequent words tend to co-occur with a lot of other words by chance. Thus, the more frequently a word is seen, the less informative its history will be, since it will reflect uninteresting co-occurrences with all sorts of words. ISA implements this by reducing the influence that the history of a context word c has on the target word t as a function of the token frequency of c (notice that the model still keeps track of the encounter with c, by adding its signature to the history of t; it is just the history of c that is weighted down). More precisely, the m weight associated with a context word c decreases as follows:

$$m_c = \frac{1}{\exp\left(\frac{Count(c)}{k_m}\right)}$$

where k_m is a parameter determining how fast the decay will be.

3 Experimental setting

3.1 The Lara corpus

The input for our experiments is provided by the Child-Directed-Speech (CDS) section of the Lara corpus (Rowland et al., 2005), a longitudinal corpus of natural conversation transcripts of a single child, Lara, between the ages of 1;9 and 3;3. Lara was the firstborn monolingual English daughter of two White university graduates and was born and brought up in Nottinghamshire, England. The corpus consists of transcripts from 122 separate recording sessions in which the child interacted with adult caretakers in spontaneous conversations. The total recording time of the corpus is of about 120 hours, representing one of the densest longitudinal corpora available. The adult CDS section we used contains about 400K tokens and about 6K types.

We are aware that the use of a single-child corpus may have a negative impact on the generalizations on semantic development that we can draw from the experiments. On the other hand, this choice has the important advantage of providing a fairly homogeneous data environment for our computational simulations. In fact, we can abstract from the intrinsic variability characterizing any multi-child corpus, and stemming from differences in the conversation settings, in the adults' grammar and lexicon, etc. Moreover, whereas we can take our experiments to constitute a (very rough) simulation of how a particular child acquires semantic representations from her specific linguistic input, it is not clear what simulations based on an "averages" of different linguistic experiences would represent.

The corpus was part-of-speech-tagged and lemmatized using the CLAN toolkit (MacWhinney, 2000). The automated output was subsequently checked and disambiguated manually, resulting in very accurate annotation. In our experiments, we use lemma-POS pairs as input to the word space models (e.g., `go-v` rather than `going`, `goes`, etc.) Thus, we make the unrealistic assumptions that the learner already solved the problem of syntactic categorization and figured out the inflectional morphology of her language. While a multi-level bootstrapping process in which the morphosyntactic and lexical properties of words are learned in parallel is probably cognitively more likely, it seems reasonable at the current stage of experimentation to fix morphosyntax and focus on semantic learning.

3.2 Model training

We experimented with three word space models: ISA, RI (our implementations in both cases) and the SVD-based technique implemented by the *Infomap* package.[1]

Parameter settings may considerably impact the performance of word space models (Sahlgren, 2006). In a stage of preliminary investigations (not reported here, and involving also other corpora) we identified a relatively small range of values for each parameter of each model that produced promising results, and we focused on it in the subsequent, more systematic exploration of the parameter space.

For all models, we used a context window of five words to the left and five words to the right of the target. For both RI and ISA, we set signature initialization parameters (determining the random assignment of 0s, +1s and -1s to signature vectors) similar to those described by Karlgren and Sahlgren (2001). For RI and SVD, we used two stop word filtering lists (removing all function words, and removing the top 30 most frequent words), as well as simulations with no stop word filtering. For RI and ISA, we used signature and history vectors of 1,800 and 2,400 dimensions (the first value, again, inspired by Karlgren and Sahlgren's work). Preliminary experiments with 300 and 900 dimensions produced poor results, especially with RI. For SVD, we used 300 dimensions only. This was in part due to technical limitations of the implementation, but 300 dimensions is also a fairly typical choice for SVD-based models such as LSA, and a value reported to produce excellent results in the literature. More importantly, in unrelated experiments SVD with 300 dimensions and function word filtering achieved state-of-the-art performance (accuracy above 90%) in the by now standard TOEFL synonym detection task (Landauer and Dumais, 1997).

After preliminary experiments showed that both models (especially ISA) benefited from a very low impact rate, the impact parameter i of RI and ISA was set to 0.003 and 0.009. Finally, k_m (the ISA parameter determining the steepness of decay of the influence of history as the token frequency of the context word increases) was set to 20 and 100 (recall that a higher k_m correspond to a less steep decay).

The parameter settings we explored were systematically crossed in a series of experiments. Moreover, for RI and ISA, given that different random initializations will lead to (slightly) different results, each experiment was repeated 10 times.

Below, we will report results for the best performing models of each type: ISA with 1,800 dimensions, i set to 0.003 and k_m set to 100; RI with 2,400 dimensions, i set to 0.003 and no stop words; SVD with 300-dimensional vectors and function words removed. However, it must be stressed that 6 out of the 8 ISA models we experimented with outperformed the best RI model (and they all outperformed the best SVD model) in the Noun AP task discussed in section 4.1. This suggests that the results we report are not overly dependent on specific parameter choices.

3.3 Evaluation method

The test set was composed of 100 nouns and 70 verbs (henceforth, Ns and Vs), selected from the most frequent words in Lara's CDS section (word frequency ranges from 684 to 33 for Ns, and from

[1] `http://infomap-nlp.sourceforge.net/`

3501 to 89 for Vs). This asymmetry in the test set mirrors the different number of V and N types that occur in the input (2828 Ns vs. 944 Vs). As a further constraint, we verified that all the words in the test set also appeared among the child's productions in the corpus. The test words were unambiguously assigned to semantic categories previously used to model early lexical development and represent plausible early semantic groupings. Semantic categories for nouns and verbs were derived by combining two methods. For nouns, we used the ontologies from the Macarthur-Bates Communicative Development Inventories (CDI).[2] All the Ns in the test set also appear in the Toddler's List in CDI. The noun semantic categories are the following (in parenthesis, we report the number of words per class and an example): ANIMALS_REAL_OR_TOY (19; *dog*), BODY_PARTS (16; *nose*), CLOTHING (5; *hat*), FOOD_AND_DRINK (13; *pizza*), FURNITURE_AND_ROOMS (8; *table*), OUTSIDE_THINGS_AND_PLACES_TO_GO (10; *house*), PEOPLE (10; *baby*), SMALL_HOUSEHOLD_ITEMS (13; *bottle*), TOYS (6; *pen*). Since categories for verbs were underspecified in the CDI, we used 12 broad verb semantic categories for event types, partly extending those in Borovsky and Elman (2006): ACTION (11; *play*), ACTION_BODY (6; *eat*), ACTION_FORCE (5; *pull*), ASPECTUAL (6; *start*), CHANGE (12; *open*), COMMUNICATION (4; *talk*), MOTION (5; *run*), PERCEPTION (6; *hear*), PSYCH (7; *remember*), SPACE (3; *stand*), TRANSFER (6; *buy*).

It is worth emphasizing that this experimental setting is much more challenging than those that are usually adopted by state-of-the-art computational simulations of word learning, as the ones reported above. For instance, the number of words in our test set is larger than the one in Borovsky and Elman (2006), and so is the number of semantic categories, both for Ns and for Vs. Conversely, the Lara corpus is much smaller than the data-sets normally used to train word space models. For instance, the best results reported by Li et al. (2000) are obtained with an input corpus which is 10 times bigger than ours.

As an evaluation measure of the model performance in the word learning task, we adopted *Average Precision* (AP), recently used by Borovsky and Elman (2006). AP evaluates how close all members of a certain category are to each other in the semantic space built by the model.

To calculate AP, for each w_i in the test set we first extracted the corresponding distributional vector v_i produced by the model. Vectors were used to calculate the pair-wise cosine between each test word, as a measure of their distance in the semantic space. Then, for each target word w_i, we built the list r_i of the other test words ranked by their decreasing cosine values with respect to w_i. The ranking r_i was used to calculate $AP(w_i)$, the Word Average Precision for w_i, with the following formula:

$$AP(w_i) = \frac{1}{|C_{w_i}|} \sum_{w_j \in C_{w_i}} \frac{n_{w_j}(C_{w_i})}{n_{w_j}}$$

where C_{w_i} is the semantic category assigned to w_i, n_{w_j} is the set of words appearing in r_i up to the rank occupied by w_j, and $n_{w_j}(C_{w_i})$ is the subset of words in n_{w_j} that belong to category C_{w_i}.

$AP(w_i)$ calculates the proportion of words that belong to the same category of w_i at each rank in r_i, and then divides this proportion by the number of words that appear in the category. AP ranges from 0 to 1: $AP(w_i) = 1$ would correspond to the ideal case in which all the closest words to w_i in r_i belonged to the same category as w_i; conversely, if all the words belonging to categories other than C_{w_i} were closer to w_i than the words in C_{w_i}, $AP(w_i)$ would approach 0. We also defined the Class AP for a certain semantic category by simply averaging over the Word $AP(w_i)$ for each word in that category:

$$AP(C_i) = \frac{\sum_{j=1}^{j=|C_i|} AP(w_j)}{|C_i|}$$

We adopted AP as a measure of the purity and cohesiveness of the semantic representations produced by the model. Words and categories for which the model is able to converge on well-formed representations should therefore have higher AP values. If we define Recall as the number of words in n_{w_j} belonging to C_{w_i} divided by the total number of words in C_{w_i}, then all the AP scores reported in our experiments correspond to 100% Recall, since the neighbourhood we used to compute $AP(w_i)$ always included all the words in C_{w_i}. This represents a very

[2] `http://www.sci.sdsu.edu/cdi/`

Nouns			
Tokens	**ISA**	**RI**	**SVD**
100k	0.321	0.317	0.243
200k	0.343	0.337	0.284
300k	0.374	0.367	0.292
400k	0.400	0.393	0.306
Verbs			
100k	0.242	0.247	0.183
200k	0.260	0.266	0.205
300k	0.261	0.266	0.218
400k	0.270	0.272	0.224

Table 1: Word AP scores for Nouns (top) and Verbs (bottom). For ISA and RI, scores are averaged across 10 iterations

stringent evaluation condition for our models, far beyond what is commonly used in the evaluation of classification and clustering algorithms.

4 Experiments and results

4.1 Word learning

Since we intended to monitor the incremental path of word learning given increasing amounts of linguistic input, AP scores were computed at four "training checkpoints" established at 100K, 200K, 300K and 400K word tokens (the final point corresponding to the whole corpus).[3] Scores were calculated independently for Ns and Vs. In Table 1, we report the AP scores obtained by the best performing models of each type , as described in section 3.2. The reported AP values refer to Word AP averaged respectively over the number of Ns and Vs in the test set. Moreover, for ISA and RI we report mean AP values across 10 repetitions of the experiment.

For Ns, both ISA and RI outperformed SVD at all learning stages. Moreover, ISA also performed significantly better than RI in the full-size input condition (400k checkpoint), as well as at the 300k checkpoint (*Welch t-test*; $df = 17, p < .05$).

One of the most striking results of these experiments was the strong *N-V asymmetry* in the Word AP scores, with the Vs performing significantly worse than the Ns. For Vs, RI appeared to have a small advantage over ISA, although it was never significant at any stage. The asymmetry is suggestive of the widely attested N-V asymmetry in child word learning. A consensus has gathered in the early word learning literature that children from several languages acquire Ns earlier and more rapidly than Vs (Gentner, 1982). An influential account explains this noun-bias as a product of language-external factors such as the different complexity of the world referents for Ns and Vs. Recently, Christiansen and Monaghan (2006) found that distributional information in English CDS was more reliable for identifying Ns than Vs. This suggests that the category-bias may also be partly driven by how good certain language-internal cues for Ns and Vs are in a given language. Likewise, distributional cues to semantics may be stronger for English Ns than for Vs. The noun-bias shown by ISA (and by the other models) could be taken to complement the results of Christiansen and Monaghan in showing that English Ns are more easily discriminable than Vs on distributionally-grounded semantic terms.

4.2 Category learning

In Table 2, we have reported the Class AP scores achieved by ISA, RI and SVD (best models) under the full-corpus training regime for the nine nominal semantic categories. Although even in this case ISA and RI generally perform better than SVD (with the only exceptions of FURNITURE_AND_ROOMS and SMALL_HOUSEHOLD_ITEMS), results show a more complex and articulated situation. With BODY_PARTS, PEOPLE, and SMALL_HOUSEHOLD_ITEMS, ISA significantly outperforms its best rival RI (*Welch t-test*; $p < .05$). For the other classes, the differences among the two models are not significant, except for CLOTHING in which RI performs significantly better than ISA. For verb semantic classes (whose analytical data are not reported here for lack of space), no significant differences exist among the three models.

Some of the lower scores in Table 2 can be explained either by the small number of class members (e.g. TOYS has only 6 items), or by the class highly heterogeneous composition (e.g. in OUTSIDE_THINGS_AND_PLACES_TO_GO we find nouns like *garden*, *flower* and *zoo*). The case of PEOPLE, for which the performance of all the three models is far below their average Class AP score (ISA = 0.35; RI = 0.35; SVD = 0.27), is instead much more surprising. In fact, PEOPLE is one of the classes

[3]The checkpoint results for SVD were obtained by training different models on increasing samples from the corpus, given the non-incremental nature of this method.

Semantic class	ISA	RI	SVD
ANIMALS_REAL_OR_TOY	0.616	0.619	0.438
BODY_PARTS	0.671	0.640	0.406
CLOTHING	0.301	0.349	0.328
FOOD_AND_DRINK	0.382	0.387	0.336
FURNITURE_AND_ROOMS	0.213	0.207	0.242
OUTSIDE_THINGS_PLACES	0.199	0.208	0.198
PEOPLE	0.221	0.213	0.201
SMALL_HOUSEHOLD_ITEMS	0.208	0.199	0.244
TOYS	0.362	0.368	0.111

Table 2: Class AP scores for Nouns. For ISA and RI, scores are averaged across 10 iterations

with the highest degree of internal coherence, being composed only of nouns unambiguously denoting human beings, such as *girl*, *man*, *grandma*, etc. The token frequency of the members in this class is also fairly high, ranging between 684 and 55 occurrences. Last but not least, in unrelated experiments we found that a SVD model trained on the British National Corpus with the same parameters as those used with Lara was able to achieve very good performances with human denoting nouns, similar to the members of our PEOPLE class.

These facts have prompted us to better investigate the reasons why with Lara none of the three models was able to converge on a satisfactory representation for the nouns belonging to the PEOPLE class. We zoomed in on this semantic class by carrying out another experiment with ISA. This model underwent 8 cycles of evaluation, in each of which the 10 words originally assigned to PEOPLE have been reclassified into one of the other nominal classes. For each cycle, AP scores were recomputed for the 10 test words. The results are reported in Figure 1 (where AP refers to the average Word AP achieved by the 10 words originally belonging to the class PEOPLE). The highest score is reached when the PEOPLE nouns are re-labeled as ANIMALS_REAL_OR_TOY (we obtained similar results in a parallel experiment with SVD). This suggests that the low score for the class PEOPLE in the original experiment was due to ISA mistaking people names for animals. What *prima facie* appeared as an error could actually turn out to be an interesting feature of the semantic space acquired by the model. The experiments show that ISA (as well as the other models) groups together animals and people Ns, as

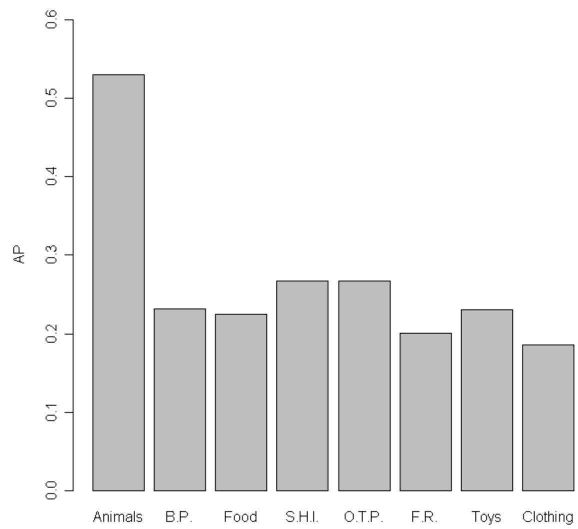

Figure 1: AP scores for Ns in PEOPLE reclassified in the other classes

it has formed a general and more underspecified semantic category that we might refer to as ANIMATE. This hypothesis is also supported by qualitative evidence. A detailed inspection of the CDS in the Lara corpus reveals that the animal nouns in the test set are mostly used by adults to refer either to toy-animals with which Lara plays or to characters in stories. In the transcripts, both types of entities display a very human-like behavior (i.e., they talk, play, etc.), as it happens to animal characters in most children's stories. Therefore, the difference between model performance and the gold standard ontology can well be taken as an interesting clue to a genuine peculiarity in children's semantic space with respect to adult-like categorization. Starting from an input in which animal and human nouns are used in similar contexts, ISA builds a semantic space in which these nouns belong to a common underspecified category, much like the world of a child in which cats and mice behave and feel like human beings.

5 Conclusion

Our main experiments show that ISA significantly outperforms state-of-the-art word space models in a learning task carried out under fairly challenging training and testing conditions. Both the incremental nature and the particular shape of the semantic representations built by ISA make it a (relatively) realistic computational model to simulate the emergence of a semantic space in early childhood.

Of course, many issues remain open. First of all, although the *Lara* corpus presents many attractive characteristics, it still contains data pertaining to a single child, whose linguistic experience may be unusual. The evaluation of the model should be extended to more CDS corpora. It will be especially interesting to run experiments in languages such as as Korean (Choi and Gopnik, 1995), where no noun-bias is attested. There, we would predict that the distributional information to semantics be less skewed in favor of nouns. All CDS corpora we are aware of are rather small, compared to the amount of linguistic input a child hears. Thus, we also plan to test the model on "artificially enlarged" corpora, composed of CDS from more than one child, plus other texts that might be plausible sources of early linguistic input, such as children's stories.

In addition, the target of the model's evaluation should not be to produce as high a performance as possible, but rather to produce performance matching that of human learners.[4] In this respect, the output of the model should be compared to what is known about human semantic knowledge at various stages, either by looking at experimental results in the acquisition literature or, more directly, by comparing the output of the model to what we can infer about the semantic generalizations made by the child from her/his linguistic production recorded in the corpus.

Finally, further studies should explore how the space constructed by ISA depends on the *order* in which sentences are presented to it. This could shed some light on the issue of how different experiential paths might lead to different semantic generalizations.

While these and many other experiments must be run to help clarifying the properties and effectiveness of ISA, we believe that the data presented here constitute a very promising beginning for this new line of research.

References

Borovsky, A. and J. Elman. 2006. Language input and semantic categories: a relation between cognition and early word learning. *Journal of Child Language*, 33: 759-790.

Burgess, C. and K. Lund. 1997. Modelling parsing constraints with high-dimensional context space. *Language and Cognitive Processes*, 12: 1-34.

Choi, S. and A. Gopnik, A. 1995. Early acquisition of verbs in Korean: a cross-linguistic study. *Journal of Child Language* 22: 497-529.

Christiansen, M.H. and P. Monaghan. 2006. Discovering verbs through multiple-cue integration. In K. Hirsh-Pasek and R.M. Golinkoff (eds.), *Action meets word: How children learn verbs*. OUP, Oxford.

Farkas, I. and P. Li. 2001. A self-organizing neural network model of the acquisition of word meaning. *Proceedings of the 4th International Conference on Cognitive Modeling*.

Gentner, D. 1982. Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In S.A. Kuczaj (ed.), *Language development, vol. 2: Language, thought and culture*. Erlbaum, Hillsdale, NJ.

Karlgren, J. and M. Sahlgren. 2001. From words to understanding. In Uesaka, Y., P. Kanerva and H. Asoh (eds.), *Foundations of real-world intelligence*, CSLI, Stanford: 294-308,

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

Li, P., C. Burgess and K. Lund. 2000. The acquisition of word meaning through global lexical co-occurrences. *Proceedings of the 31st Child Language Research Forum*: 167-178.

Li, P., I. Farkas and B. MacWhinney. 2004. Early lexical acquisition in a self-organizing neural network. *Neural Networks*, 17(8-9): 1345-1362.

Manning Ch. and H. Schütze. 1999. *Foundations of statistical natural language processing* The MIT Press, Cambridge, MASS.

MacWhinney, B. 2000. *The CHILDES project: Tools for analyzing talk (3d edition).* Erlbaum, Mahwah, NJ.

Rowland, C., J. Pine, E. Lieven and A. Theakston. 2005. The incidence of error in young children's wh-questions. *Journal of Speech, Language and Hearing Research*, 48(2): 384-404.

Sahlgren, M. 2006. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces.* Ph.D. dissertation, Department of Linguistics, Stockholm University.

[4]We thank an anonymous reviewer for this note