Climbing the path to grammar: a maximum entropy model of subject/object learning

Felice Dell'Orletta
Dept. of Computer Science
University of Pisa
Largo Pontecorvo 3
56100 Pisa (Italy)

Alessandro Lenci
Dept. of Linguistics
University of Pisa
Via Santa Maria 36
56100 Pisa (Italy)

Simonetta Montemagni
ILC-CNR
Area della Ricerca
Via Moruzzi 1
56100 Pisa (Italy)

Vito Pirrelli
ILC-CNR
Area della Ricerca
Via Moruzzi 1
56100 Pisa (Italy)

{felice.dellorletta, alessandro.lenci, simonetta.montemagni, vito.pirrelli}@ilc.cnr.it

Abstract

In this paper, we discuss an application of Maximum Entropy to modeling the acquisition of subject and object processing in Italian. The model is able to learn from corpus data a set of experimentally and theoretically well-motivated linguistic constraints, as well as their relative salience in Italian grammar development and processing. The model is also shown to acquire robust syntactic generalizations by relying on the evidence provided by a small number of high token frequency verbs only. These results are consistent with current research focusing on the role of high frequency verbs in allowing children to converge on the most salient constraints in the grammar.

1 Introduction

Current research in language learning supports the view that developing grammatical competence involve mastering and integrating multiple, parallel, probabilistic constraints defined over different types of linguistic (and non linguistic) information (Seidenberg and MacDonald 1999, MacWhinney 2004). This is particularly clear when we focus on the core of grammatical development, namely the ability to properly identify syntactic relations. Psycholinguistic evidence shows that children learn to identify sentence subjects and direct objects by combining various types of probabilistic cues, such as word order, noun animacy, definiteness, agreement, etc. The relative prominence of each of these cues during the development of a child's syntactic competence can considerably vary cross-linguistically, mirroring their relative salience in the adult grammar system (cf. Bates *et al.* 1984).

If grammatical constraints are inherently probabilistic (Manning 2003), the path through which the child acquires adult grammar competence can be viewed as the process of building a stochastic model out of the linguistic input. Consistently with "usage-based" approaches to language acquisition (cf. Tomasello, 2000) grammatical constraints would thus emerge from language use thanks to the child's ability to keep track of statistical regularities in linguistic cues. In turn, this raises the issue of how children are able to exploit the statistical distribution of cues in the linguistic input. Various types of cross-linguistic evidence converge on the hypothesis that children are actually able to take great advantage of the highly skewed distribution of naturalistic language data. Goldberg *et al.* (2004), Matthews *et al.* (2003), Ninio (1999) among the others argue that verbs with high token frequency in the input have a facilitatory effect in allowing children to derive robust syntactic generalizations even from surprisingly minimal input. According to this model, syntactic learning is driven by a small pool of verbs occurring with the highest token frequency: they approximately correspond to so-called "light verbs" such as English *go*, *give*, *want* etc. These verbs would act as "cata-

Proceedings of the Second Workshop on Psychocomputational Models of Human Language Acquisition, pages 72–81,
Ann Arbor, June 2005.

lysts" in allowing children to converge on the most salient grammar constraints of the language they are acquiring.

In computational linguistics, *Maximum Entropy* models have proven to be robust statistical learning algorithms that perform well in a number of processing tasks (cf. Ratnaparkhi 1998). In this paper, we discuss successful application of a Maximum Entropy (*ME*) model to the processing of Italian syntactic relations. We believe that this discussion is of general interest for two basic reasons. First, the model is able to learn, from corpus data, a set of experimentally and theoretically well-motivated linguistic constraints, as well as their relative salience in the processing of Italian. This suggests that it is possible for a child to bootstrap and use this type of knowledge on the basis of a specific distribution of real language data, a conclusion that bears on the question of the role and type of innate inductive biases. Secondly, the model is also shown to acquire robust syntactic generalizations by relying on the evidence provided by a small number of high token frequency verbs only. With some qualifications, this evidence sheds light on the interaction between highly skewed language data distributions and language maturation. Robust grammar generalizations emerge on the basis of exposure to early, statistically stable and lexically underspecified evidence, thus providing a reliable backbone to children's syntactic development and later lexical organization.

In the following section we first broach the general problem of parsing subjects and objects in Italian. Section 3 describes an *ME* model of the problem. Section 4 and 5 are devoted to a detailed empirical analysis of the interaction of different feature configurations and of the interplay between verb token frequency and relevant generalizations. Conclusions are drawn in the final discussion.

2 Subjects and Objects in Italian

Children that learn how to process subjects and objects in Italian are confronted with a twofold challenge: i) the relatively free order of Italian sentence constituents and ii) the possible absence of an overt subject. The existence of a preferred Subject Verb Object (*SVO*) order in Italian main clauses does not rule out all other possible permutations of these units: in fact, they are all attested, albeit with considerable differences in distribution and degree of markedness (Bartolini *et al.* 2004).[1] Moreover, because of pro-drop, an Italian Verb Noun (*VN*) sequence can either be interpreted as a *VO* construction with subject omission (e.g. *ha dichiarato guerra* '(he) declared war') or as an instance of postverbal subject (*VS*, e.g. *ha dichiarato Giovanni* 'John declared'). Symmetrically, an *NV* sequence is potentially ambiguous between *SV* and *OV*: compare *il bambino ha mangiato* 'the child ate' with *il gelato ha mangiato* 'the ice-cream, (he) ate'.

These grammatical facts are in keeping with what we know about Italian children's parsing strategies. Bates *et al.* (1984) show that while, in English, word order is by and large the most effective cue for subject-object identification (henceforth *SOI*) both in syntactic processing and during the child's syntactic development, the same cue plays second fiddle in Italian. Bates and colleagues bring empirical evidence supporting the hypothesis that Italian children show extreme reliance on *NV* agreement and, secondly, on noun animacy, rather than word order. They conclude that the following syntactic constraints dominance hierarchy is operative in Italian: agreement > animacy > word order.

The fact that animacy can reliably be resorted to in Italian *SOI* receives indirect confirmation from corpus data. We looked at the distribution of animate subjects and objects in the Italian Syntactic Semantic Treebank (ISST, Montemagni *et al.*, 2003), a 300,000 tokens syntactically annotated corpus, including articles from contemporary Italian newspapers and periodicals covering a broad variety of topics. Subjects and objects in ISST were automatically annotated for animacy using the SIMPLE Italian computational lexicon (Lenci *et al.* 2000) as a background semantic resource. The annotation was then checked manually. Corpus analysis highlights a strong asymmetry in the distribution of animate nouns in subject and object roles: over 56.6% of ISST subjects are animate (out of a total number of 12,646), while only the 11.1% of objects are animate (out of a total number of 5,559). Such an overwhelming preference for inanimate objects in adult language data makes animacy play a very important role in *SOI*, both as a key developmental factor in the bootstrapping of the syntax-semantics mapping and as a reliable

[1] In the present paper we restrict ourselves to the case of declarative main clauses.

processing cue, consistently with psycholinguistic data.

On the other hand, the distribution of word order configurations in the same corpus shows another interesting asymmetry. *NV* sequences receive an *SV* interpretation in 95.6% of the cases, and an object interpretation in the remaining 4.4% (most of which are clitic and relative pronouns, whose preverbal position is grammatically constrained). The situation is quite different when we turn to *VN* sequences, where verb-object pairs represent 73.4% of the cases, with verb-subject pairs representing the remaining 26.6%. We infer that – at least in standard written Italian – *VS* is a much more consistently used construction than *OV*, and that the role of word order in Italian parsing is not a marginal one across the board, but rather *relative* to *VN* contexts only. In *NV* constructions there is a strong preference for a subject interpretation, and this suggests a more dynamic dominance hierarchy of Italian syntactic constraints than the one provided above.

As for agreement, it represents conclusive evidence for *SOI* only when a nominal constituent and a verb do not agree in number and/or person (as in *leggono il libro* '(they) read the book'). On the contrary, when noun and verb share the same person and number the impact of agreement on *SOI* is neutralised, as in *il bambino legge il libro* 'the child reads the book' or in *ha dichiarato il presidente* 'the president declared'. Although this ambiguity arises in specific contexts (i.e. when the verb is used in the third person singular or plural and the subject/object candidate agrees with it), it is interesting to note that in ISST: third person verb forms cover 95.6% of all finite verb forms; and, more interestingly for our present concerns, 87.9% of all *VN* and *NV* pairs involving a third person verb form contains an agreeing noun. From this we conclude that the contribution of agreement to our problem is fairly limited, as *lack* of agreement shows up only in a limited number of contexts.

All in all, corpus data lend support to the idea that in Italian *SOI* is governed by a complex interplay of probabilistic constraints of a different nature (morpho-syntactic, semantic, word order etc.). Moreover, distributional asymmetries in language data seem to provide a fairly reliable statistical basis upon which relevant probabilistic constraints can be bootstrapped and combined consistently. In the following section we shall present a ME model of how constraints and their interaction can be bootstrapped from language data.

3 A Maximum Entropy model of SOI

The Maximum Entropy (ME) framework offers a mathematically sound way to build a probabilistic model for *SOI*, which combines different linguistic cues. Given a linguistic context c and an outcome $a \in A$ that depends on c, in the ME framework the conditional probability distribution $p(a|c)$ is estimated on the basis of the assumption that no *a priori* constraints must be met other than those related to a set of features $f_j(a,c)$ of c, whose distribution is derived from the training data. It can be proven that the probability distribution p satisfying the above assumption is the one with the highest entropy, is unique and has the following exponential form (Berger *et al.* 1996):

$$(1) \qquad p(a \mid c) = \frac{1}{Z(c)} \prod_{j=1}^{k} \alpha_j^{f_j(a,c)}$$

where $Z(c)$ is a normalization factor, $f_j(a,c)$ are the values of k features of the pair (a,c) and correspond to the linguistic cues of c that are relevant to predict the outcome a. Features are extracted from the training data and define the constraints that the probabilistic model p must satisfy. The parameters of the distribution $\alpha_1, \ldots, \alpha_k$ correspond to *weights* associated with the features, and determine the relevance of each feature in the overall model. In the experiments reported below feature weights have been estimated with the Generative Iterative Scaling (GIS) algorithm implemented in the AMIS software (Miyao and Tsujii 2002).

We model *SOI* as the task of predicting the correct syntactic function $f \in \{subject, object\}$ of a noun occurring in a given syntactic context s. This is equivalent to build the conditional probability distribution $p(f|s)$ of having a syntactic function f in a syntactic context s. Adopting the ME approach, the distribution p can be rewritten in the parametric form of (1), with features corresponding to the linguistic contextual cues relevant to *SOI*. The context s is a pair $\langle v_s, n_s \rangle$, where v_s is the verbal head and n_s its nominal dependent in s. This notion of s departs from more traditional ways of describing an *SOI* context as a *triple* of one verb and two nouns in a certain syntactic configuration (e.g, SOV or VOS, etc.). In fact, we assume that *SOI* can be stated in terms of the more

local task of establishing the grammatical function of a noun *n* observed in a verb-noun pair. This simplifying assumption is consistent with the claim in MacWhinney *et al.* (1984) that *SVO* word order is actually derivative from *SV* and *VO* local patterns and downplays the role of the transitive complex construction in sentence processing. Evidence in favour of this hypothesis also comes from corpus data: in ISST, there are 4,072 complete subject-verb-object-configurations, a small number if compared to the 11,584 verb tokens appearing with either a subject or an object only. Due to the comparative sparseness of canonical *SVO* constructions in Italian, it seems more reasonable to assume that children should pay a great deal of attention to both *SV* and *VO* units as cues in sentence perception (Matthews *et al.* 2004). Reconstruction of the whole lexical *SVO* pattern can accordingly be seen as the end point of an acquisition process whereby smaller units are re-analyzed as being part of more comprehensive constructions. This hypothesis is more in line with a *distributed* view of canonical constructions as derivative of more basic local positional patterns, working together to yield more complex and abstract constructions. Last but not least, assuming verb-noun pairs as the relevant context for *SOI* allows us to simultaneously model the interaction of word order variation with pro-drop in Italian.

4 Feature selection

The most important part of any ME model is the selection of the context features whose weights are to be estimated from data distributions. Our feature selection strategy is grounded on the main assumption that *features should correspond to linguistically and psycholinguistically well-motivated contextual cues*. This allows us to evaluate the probabilistic model also with respect to its ability to replicate psycholinguistic experimental results and to be consistent with linguistic generalizations.

Features are binary functions $f_{k_i,f}(f,s)$, which test whether a certain *cue* k_i for the function f occurs in the context s. For our ME model of *SOI*, we have selected the following types of features:

Word order tests the position of the noun wrt the verb, for instance:

(2)

$$f_{post,subj}(subj,s) = \begin{cases} 1 & if\ noun_s.pos = post \\ 0 & otherwise \end{cases}$$

Animacy tests whether the noun in s is *animate* or *inanimate* (cf. §.2). The centrality of this cue in Italian is widely supported by psycholinguistic evidence. Another source of converging evidence comes from functional and typological linguistic research. For instance, Aissen (2003) argues for the universal value of the following hierarchy representing the relative markedness of the associations between grammatical functions and animacy degrees (with each item in these scale been less marked than the elements to its right):

Animacy Markedness Hierarchy
Subj/Human > Subj/Animate > Subj/Inanimate
Obj/Inanimate > Obj/Animate > Obj/Human

Markedness hierarchies have also been interpreted as probabilistic constraints estimated form corpus data (Bresnan *et al.* 2001, Øvrelid 2004). In our ME model we have used a reduced version of the animacy markedness hierarchy in which human and animate nouns have been both subsumed under the general class *animate*.

Definiteness tests the degree of "referentiality" of the noun in a context pair s. Like for animacy, definiteness has been claimed to be associated with grammatical functions, giving rise to the following universal markedness hierarchy Aissen (2003):

Definiteness Markedness Hierarchy
Subj/Pro > Subj/Name > Subj/Def > Subj/Indef
Obj/Indef > Obj/Def > Obj/Name > Obj/Pro

According to this hierarchy, subjects with a low degree of definiteness are more marked than subjects with a high degree of definiteness (for objects the reverse pattern holds). Given the importance assigned to the definiteness markedness hierarchy in current linguistic research, we have included the definiteness cue in the ME model. It is worth remarking that, unlike animacy, in psycholinguistic experiments definiteness has not been assigned any effective role in *SOI*. This makes testing this cue in a computational model even more interesting, as a way to evaluate its effective contribution to Italian *SOI*. In our experiments, we have used a "compact" version of the definiteness scale: the definiteness cue tests whether the noun in the context

pair i) is a name or a pronoun ii) has a definite article iii), has an indefinite article or iv) is a "bare" noun (i.e. with no article). It is worth saying that "bare" nouns are usually placed at the bottom end of the definiteness scale.

The three types of features above only refer to nominal cues in the context pairs. Nevertheless, specific lexical properties of the verb can also be resorted to in *SOI*. The probability for n_s to be subject or object may also depend on the specific lexical preferences of v_s. To take this lexical factor into account, we add a set of *lexical cues* to the three general feature types above. Lexical cues test animacy with respect to a specific verb v_k:

(3)

$$f_{anim,v_k,subj}(subj,s) = \begin{cases} 1 & if \quad v_k = v_s \wedge n_s = anim \\ 0 & otherwise \end{cases}$$

Lexical features provide evidence of the propensity of a given verb to have an animate (inanimate) subject or object. In fact, the verb argument structure and thematic properties may well influence the possible distribution of animate (inanimate) subjects and objects, thus overriding more general tendencies. By including lexical cues, we are thus able to test the interplay of lexical constraints with general grammatical ones.

Note that in our ME model we have not included agreement as a feature, in spite of its prominent role in Italian. The fact that agreement is often inconclusive for *SOI* (§.2) suggests that children must also acquire the ability to deal with the interplay of various concurrent constraints, none of which is singularly sufficient for the task completion this type of competence. It is exactly this area of syntactic competence that we wanted to explore with the experiments reported below (cf. MacWhinney *et al.* 1984, who similarly abstract from the dominant role of case in German *SOI*).

5 Testing feature configurations for *SOI*

The ME model for Italian *SOI* has been trained on 18,205 verb-subject/object pairs extracted from ISST. The training set was obtained by extracting all verb-subject and verb-object dependencies headed by an active verb occurring in a finite verbal construction and by excluding all cases where the position of the nominal constituent was grammatically constrained (e.g. clitic objects, relative clauses). Two different feature configurations have been used for training:

- *non-lexical feature configuration* (*NLC*), including only general features acting as global constraints: namely word order, noun animacy and noun definiteness;
- *lexical feature configuration* (*LC*), including word order, noun animacy and definiteness, and information about the verb head.

The test corpus consists of 645 verb-noun pairs extracted from contexts where agreement happens to be neutralized. Of them, 446 contained a subject (either pre- or post-verbal) and 199 contained an object (either pre- or post-verbal). The two feature configurations were evaluated by calculating the percentage of correctly assigned relations over the total number of test pairs (accuracy). As our model always assigns one syntactic relation to each test pair, accuracy equals both standard precision and recall. Finally, we have assumed a baseline score of 69%, corresponding to the result yielded by a dumb model assigning to each test pair the most frequent relation in the training corpus, i.e. subject.

5.1 Non-lexical feature configuration

Our first experiment was carried out with *NLC*. The accuracy on the test corpus is 91.5%; most errors (i.e. 96.4%) relate to the postverbal position, with 44 mistaken subjects (42 inanimate) and 9 mistaken objects (all animate). The score was confirmed by a 10-fold cross-validation on the whole training set (89.3% accuracy).

A further way to evaluate the goodness of the model is by inspecting the weights associated with feature values (Table 1).

	Subj	**Obj**
Preverbal	1,34E+00	2,10E-02
Postverbal	5,21E-01	1,47E+00
Anim	1,28E+00	3,34E-01
Inanim	8,60E-01	1,21E+00
PronName	1,22E+00	5,75E-01
DefArt	1,05E+00	1,00E+00
IndefArt	8,33E-01	1,16E+00
NoArticle	9,46E-01	1,07E+00

Table 1 – *Feature value weights in NLC*

The grey cells in Table 1 highlight the preference of each feature value for either subject or object identification: e.g. preverbal subjects are strongly preferred over preverbal objects; animate subjects

are preferred over animate objects, etc. Interestingly, if we rank the *Anim* and *Inanim* values for subjects and objects, we can observe that they distribute consistently with the *Animacy Markedness Hierarchy* reported in §.4: *Subj*/*Anim* > *Subj*/*Inanim* and *Obj*/*Inanim* > *Obj*/*Anim*. Similarly, by ranking the values of the definiteness features in the *Subj* column by decreasing weight values we obtain the following ordering: *PronName* > *DefArt* > *IndefArt* > *NoArt*, which nicely fits in with the *Definiteness Markedness Hierarchy* in §.4. The so-called "markedness reversal" is observed if we focus on the values for the same features in the *Obj* column: the *PronName* feature represents the most marked option, followed by *DefArt*. The only exception is represented by the relative ordering of *IndefArt* and *NoArt* which however show very close values.

Evaluating feature salience

In order to evaluate the most reliable cues in Italian *SOI*, we have analysed the model predictions for different bundles of feature values. For each of the 16 different bundles (b) attested in the data, we have estimated $p(subj|b)$ and $p(obj|b)$:

b	*p*(*subj*|b)	*p*(*obj*|b)
Pre Anim IndefArt	0,994	0,006
Pre Anim DefArt	0,996	0,004
Pre Anim NoArt	0,995	0,005
Pre Anim PronName	0,998	0,002
Pre Inanim IndefArt	0,970	0,030
Pre Inanim DefArt	0,979	0,021
Pre Inanim NoArt	0,976	0,024
Pre Inanim PronName	0,990	0,010
Post Anim IndefArt	0,495	0,505
Post Anim DefArt	0,589	0,411
Post Anim NoArt	0,546	0,454
Post Anim PronName	0,743	0,257
Post Inanim IndefArt	0,153	0,847
Post Inanim DefArt	0,209	0,791
Post Inanim NoArt	0,182	0,818
Post Inanim PronName	0,348	0,652

Table 2 – *Subj/obj probabilities by different bundles*

The model shows a neat preference for subject when the noun is preverbal. Instead, when the noun is postverbal, function assignment is *de facto* decided by the noun animacy. Conversely, definiteness features have a much more secondary role: they can re-enforce (or weaken) the preference expressed by animacy, but they do not have the strength to determine *SOI*.

The relative salience of the different constraints acting on *SOI* can also be inferred by comparing the weights associated with individual feature values. For instance, Goldwater and Johnson (2003) show that ME can be successfully applied to learn constraint rankings in Optimality Theory, by assuming the parameter weights a_1, …, a_k as the ranking values of the constraints. The following table lists the 16 general constraints of the model by increasing weight values:

Feature	Weight
Preverbal_Obj	2,10E-02
Anim_Obj	3,34E-01
Postverbal_Subj	5,21E-01
ProName_Obj	5,75E-01
IndefArt_Subj	8,33E-01
Inanim_Subj	8,60E-01
NoArticle_Subj	9,46E-01
ArtDef_Obj	1,00E+00
DefArt_Subj	1,05E+00
NoArticle_Obj	1,07E+00
IndefArt_Obj	1,16E+00
Inanim_Obj	1,21E+00
PronName_Subj	1,22E+00
Anim_Subj	1,28E+00
Preverbal_Subj	1,34E+00
Postverbal_Obj	1,47E+00

Table 3 – *Constraint weights ranking*

The rankings in Table 3 can be used to derive the relative salience of each constraint. Lower ranked constraints correspond to more marked syntactic configurations that are then disfavoured in *SOI*. Notice that the two animacy constraints *Anim_Obj* and *Anim_Subj* are respectively placed near the bottom and the top end of the scale. Notwithstanding the low position of *Postverbal_Subj*, animacy is thus able to override the word order constraint and to produce a strong tendency to identify animate nouns as subjects, even when they appear in postverbal position (cf. Table 2 above). The constraint ranking thus confirms the interplay between animacy and word order in Italian, with the former playing a decisive role in assigning the syntactic function of postverbal nouns. On the other hand,

the constraints involving noun definiteness occupy a more intermediate position in the general ranking, with very close values. This is again consistent with the less decisive role of this feature type in *SOI*, as shown above.

5.2 Lexical feature configuration

In this experiment the general features reported in Table 1 have been integrated with 4,316 verb-specific features as the ones exemplified below for the verb *dire* 'say':

*dire*_animSog	1.228213e+00
*dire*_noanimSog	7.028484e-01
*dire*_animOgg	3.645964e-01
*dire*_noanimOgg	1.321887e+00

whose associated weights show the strong preference of this verb to take animate subjects as opposed to inanimate ones as well as a preference for inanimate objects with respect to animate ones. The results achieved with *LC* on the test corpus show a significant improvement with respect to those obtained with *NLC*: the accuracy is now 95.5%, with a 4% improvement, confirmed by a 10-fold cross-validation (94.9%). Also in this case, most of the errors relate to the postverbal position (i.e. 27 out of 29), partitioned into 26 mistaken subjects and 1 mistaken object. Lexical features have been resorted to to solve most of the *NLC* errors (i.e. 34 out of 55). It is interesting to note however that lexical features can also be misleading. The *LC* results include 8 new errors, suggesting that lexical features do not always provide conclusive evidence: in fact, in 185 cases out of 645 test *VN* pairs (i.e. 28.7% of the cases) general features are preferred over lexical ones. It is also worth mentioning that the ranking of general animacy and definiteness features in *LC* actually fits in with the respective markedness hierarchies even with a better approximation than the one produced by *NLC*. Finally, the relative prominence of the different global features confirms the trend in Table 2, with word order being predominant in preverbal position and animacy playing a major role with postverbal nouns.

Both feature configurations of the ME model thus appear to comply with linguistic and psycholinguistic generalizations on *SOI*. On the linguistic side, the constraints learnt by the model are consistent with universal markedness hierarchies for grammatical relations. Secondly, the prominence of the various constraints in the model fits in well with psycholinguistic data. Consistently with the results in Bates *et al.* (1984), the model confirms the great impact of noun animacy in Italian, although in this case its key role seems to be more directly limited to the postverbal position. Conversely, the preverbal position is by itself a very strong cue for subject interpretation.

6 High frequency verbs and *SOI*

Frequency is known to play a major influence in language learning. In morphology, for example, highly frequent lexical items tend to be shorter forms, more readily accessible in the mental lexicon, independently stored as whole items (Stemberger and MacWhinney 1986) and fairly resistant to morphological overgeneralization through time, thus establishing a correlation between irregular inflected forms and frequency. Frequency has also been assigned a key role in the acquisition of syntactic constructions. In fact, Goldberg (1998) and Ninio (1999) have independently argued for the existence of a causal relation between early exposure to highly frequent light verbs and acquisition of abstract syntax-semantics mappings (constructions). Light verbs such as *want, put* and *go* tend to be very frequent, because they are applicable in a wider range of contexts and are learned and used at an early language maturation stage The main idea is that children's early use of these high frequency verbs is conducive to the acquisition of abstract constructional properties generalizing over particular instances.

Goldberg *et al*. (2004) motivate this hypothesis by observing that light verbs have high input frequency in the child's developmental environment and, at the same time, exhibit a low degree of semantic specialization. Hence, she argues, it takes a little abstraction step for a child to jump from actual instances of use of light verbs to the syntax-semantics association of their underlying construction. On the other hand, Ninio (1999) grounds the facilitatory role of highly frequent verbs on their being "pathbreaking" *prototypes* of the construction they instantiate, since they are the best models of the relevant combinatorial and semantic properties of their construction in a relatively undiluted fashion. However, in the case of light verb constructions, the correlation between high frequency

and construction prototypicality and extension is tenuous. In fact, it is difficult to argue that frequent light verbs such as *see*, *want* or *do* exhibit a high degree of both semantic and constructional transitivity (Goldberg *et al.* 2004). This is reminiscent of the morphological behaviour of very frequent word forms in inflectional languages, as most of these forms are highly fused and show a general tendency towards irregular inflection and low morphological prototypicality. Furthermore, it is difficult to reconcile the "pathbreaking" view with the observation that frequently observed linguistic units are memorized in full, as unanalyzed wholes.

6.1 Testing the role of frequency

To address these open issues and put the alleged "pathbreaking" role of light verbs to the challenging test of a probabilistic model, we carried out a second battery of experiments to learn the general, non-lexical constraints from two training corpora of roughly equivalent size where overall type and token verb frequencies were controlled for. Both corpora are a subset of the original training set:

1. *skewed frequency corpus* (SF) – it includes 5,261 context pairs, obtained by selecting 15 verbs occurring more than 100 times in ISST (figures in parentheses give their token frequency): *essere* 'be' (2406), *avere* 'have' (708), *fare* 'do, make' (527), *dire* 'say, tell' (275), *dare* 'give' (173), *vedere* 'see' (134), *andare* 'go' (126), *sembrare* 'seem' (124), *cercare* 'try' (122), *mettere* 'put' (122), *portare* 'take' (121), *trovare* 'find' (112), *volere* 'want' (105), *lasciare* 'leave' (105), *riuscire* 'manage' (101). It is worth noticing that this set includes typical "pathbreaking" verbs;
2. *balanced frequency corpus* (BF) – this corpus includes 5,373 context pairs selected in such a way to ensure that every verb type in the original training set is attested in BF and occurs at most 6 times. For verbs occurring with a higher frequency, the pairs to be included in BF have been randomly selected.

Thus SF and BF represent two opposite training situations: SF contains few types with very high token frequencies, while BF contains a high number of verb types (i.e. 1457), with very low and uniform token frequency. These training sets resemble the structure of linguistic input used by Goldberg *et al.* (2004) for their experiments. In that case, one group of subjects was exposed to linguistic inputs in which some verbs occurred with a much higher frequency than the others; a second group of subjects was instead exposed to linguistic stimuli in which every verb occurred with roughly equal frequency. Therefore, by training our ME model on SF and BF we are able to evaluate the effective role of high token frequency verbs in driving syntactic learning.

The ME model with the general features only (i.e. *NLC*) was first trained on SF, and then tested on the 645-pair corpus in §.5, showing a 90% accuracy. The same ME model was then trained on BF, and then tested on the 645-pair corpus, scoring a 87% accuracy. The ME model trained on the skewed frequency data thus outperforms the model trained on BF in a statistically significant way ($?^2 = 4.97$; $a=0.05$; $p\text{-}value = 0.025$).

By using a training set formed only by the verbs with the highest token frequency, the model has thus been able to acquire robust syntactic constraints for *SOI*. Once these constraints have been applied to unseen events, the model has achieved a performance comparable to the one of the general models in §.5. This is somehow even more significant if we consider that the training set was now formed by less than one-third of the pairs on which the models in §.5 were trained. Data quantity aside, the most relevant fact is that it is the way verb frequencies are distributed to determine the learning path, with a significant positive effect produced by high token frequency verbs. In the model trained on SF, feature ranking is also governed by markedness relations, and the relative prominence of the various constraints is utterly similar to the one discussed in §.5. In other terms, the results of this experiment prove that frequent verbs are actually able to act as "*catalysts*" of the syntactic acquisition process. It is possible for children to converge on the correct generalizations governing *SOI* in Italian, just by relying on the linguistic evidence provided by the most frequent verbs.

This view suggests a way out of the apparent paradox of the "pathbreaking" hypothesis: highly frequent verbs can be assumed to provide stable and consistent multiple probabilistic cues for the assignment of subject/object relations. The existence of positional patterns that occur with high token frequency may well provide a deeply entrenched and highly salient set of distributional cues that act as probabilistic constraints on constructional generalizations. We hypothesize that similar constructions of other less frequent verbs

are processed, for lack of more specific overriding information, in the light of these constraints. Since processing is the result of a "conspiracy" of distributed constraints, "pathbreaking" prototypes need not be real construction exemplars but highly schematic patterns. We proved that highly frequent local positional patterns offer the right sort of constraint conspiracy.

7 General discussion

It appears that the distributional evidence of high frequency light verbs may well provide a solid cognitive anchor for sweeping perceptual generalizations on the syntax-semantics mapping. These generalizations are *local*, in that they involve positional *NV* and *VN* pairs only, and are *perceptual* as they address the issue of identifying appropriate syntactic relations by relying on perceptual features of linguistic contexts, such as position, animacy, etc. On the basis of these findings, one can reasonably argue that complex lexical constructions (in the sense of Goldberg 1998) are built upon these local patterns, by combining them in those contexts where the presence of a particular verb licenses such a combination.

The two feature configurations discussed in §.5 (i.e. *NLC* and *LC*) can thus be viewed as two successive steps along the path that leads towards the emergence of complex, lexically-driven constructions. This can actually be modeled as the incremental process of adding more and more lexical constraints to early lexicon-free generalizations (based on word order, animacy, definiteness etc.). As a result of such additional constraints, the presence of an intransitive verb may completely rule out the object interpretation of a *VN* pattern, flying in the face of a general bias towards viewing *VN* as a transitive pattern. This picture is compatible with the well-known observation that constructions are used rather conservatively by children at early stages of language maturation (Tomasello 2000). In fact, if early generalizations are mainly perceptual and local, we do not expect them to be used in production, at least until the child reaches a stage where they are combined into bigger lexically-driven constructions.

ME has proven to be a sound computational learning framework to simulate the interplay of complex probabilistic constraints in language. Our experiments confirm linguistic generalizations and psycholinguistic data for subjects and objects in Italian, while raising new interesting issues at the same time. This is the case of the role of definiteness in *SOI*. In fact, the model features neatly reproduce the definiteness markedness hierarchy, but definiteness does not appear to be really influential for subject and object processing. Various hypotheses are compatible with such results, including that definiteness is not a cue on which speakers rely for *SOI* in Italian. Another more interesting possibility is that definiteness constraints may indeed play a decisive role when the learner is asked to assign subject and object relations in the context of a more complex construction than a simple *NV* pair. Suppose that both nouns of a noun-noun-verb triple are amenable to a subject interpretation, but that one of them is a more likely subject than the other due to its being part of a definite noun phrase. Then, it is reasonable to expect that the model would select the definite noun phrase as the subject in the triple and opt for an object interpretation of the other candidate noun phrase.

As part of our future work, we plan to train the ME model on a more realistic corpus of parental input to Italian children, available in the CHILDES database (MacWhinney, 2000: http://childes.psy.cmu.edu/data/Romance/Italian). In fact, there is converging evidence that the use of particular constructions in parental speech is largely dominated by the use of each construction with one specific, highly frequent verb (e.g. *go* for the intransitive construction). The same trends noted in mother's speech to children are mirrored in children's early speech (Goldberg *et al*., 2004). Quochi (in preparation) reports a similar distributional pattern for the caused motion and intransitive motion verbs in two Italian CHILDES corpora (named "Italian-Antelmi" and "Italian-Calambrone"). If these findings are confirmed, the high accuracy of our ME model trained on the skewed frequency corpus (SF) allows us to expect an equally high accuracy when training the model on evidence from Italian parental speech.

This brings us to another related point: lack of correction/supervision in parental input. Since our ME model heavily relies on previously classified noun-verb pairs, we can legitimately wonder how easily it can be extended to simulate child language learning in an unsupervised mode. In fact, it should be appreciated that, in our experiments, compartively little rests on supervised classification. Iden-

tification of the contextually-relevant subject is, for lack of explicit morphosyntactic clues such as agreement and diathesis, simply a matter of guessing the more likely *agent* of the action expressed by the verb on the basis of semantic and pragmatic features such as animacy, definiteness and noun position to the verb. *Mutatis mutandis*, the same holds for object identification. It is then highly likely that salient evidence for the correct subject/object classification comes to the child from direct observation of the situation described by a sentence. It is such systematic coupling of linguistic evidence from the sentence with perceptual evidence of the situation described by the sentence that can assist the child in developing interface notions such as subject, object and the like.

References

Aissen J., 2003. Differential object marking: iconicity vs. economy. *Natural Language and Linguistic Theory*, 21: 435-483.

Bartolini R., Lenci A., Montemagni S., Pirrelli V., 2004. Hybrid constraints for robust parsing: First experiments and evaluation. *LREC2004*: 859-862.

Bates E., MacWhinney B., Caselli C., Devescovi A., Natale F., Venza V., 1984. A crosslinguistic study of the development of sentence interpretation strategies. *Child Development*, 55: 341-354.

Berger A., Della Pietra S., Della Pietra V., 1996. A maximum entropy approach to natural language processing. *Computational Linguistics* 22(1): 39-71

Bresnan J., Dingare D., Manning C. D., 2001. Soft constraints mirror hard constraints: voice and person in English and Lummi. *Proceedings of the LFG01 Conference*, Hong Kong: 13-32.

Goldberg A. E., 1998. The emergence of the semantics of argument structure constructions. In B. MacWhinney (ed.), The *Emergence of Language*. Lawrence Erlbaum Associates, Hillsdale, N. J.: 197-212.

Goldberg A. E., Casenhiser D., Sethuraman N., 2004. Learning argument structure generalizations, *Cognitive Linguistics*.

Goldwater S., Johnson M. 2003. Learning OT Constraint Rankings Using a Maximum Entropy Model. In Spenader J., Eriksson A., Dahl Ö. (eds.), *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*. April 26-27, 2003, Stockholm University: 111-120.

Lenci A. *et al.*, 2000. SIMPLE: A General Framework for the Development of Multilingual Lexicons. *International Journal of Lexicography*, 13 (4): 249-263.

Manning C. D., 2003. Probabilistic syntax. In R. Bod, J. Hay, S. Jannedy (eds), *Probabilistic Linguistics*, MIT Press, Cambridge MA: 289-341.

MacWhinney, B., 2000. The CHILDES project: Tools for analyzing talk. Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates

MacWhinney B., Bates E., Kliegl R., 1984. Cue validity and sentence interpretation in English, German, and Italian. *Journal of Verbal Learning and Verbal Behavior*, 23: 127-150.

MacWhinney B., 2004. A unified model of language acquisition. In J. Kroll & A. De Groot (eds.), *Handbook of bilingualism: Psycholinguistic approaches*, Oxford University Press, Oxford.

Matthews D., Lieven E., Theakston A., Tomasello M., in press, The role of frequency in the acquisition of English word order, *Cognitive Development*.

Miyao Y., Tsujii J., 2002. Maximum entropy estimation for feature forests. *Proc. HLT2002.*

Montemagni S. *et al.* 2003. Building the Italian syntactic-semantic treebank. In Abeillé A. (ed.) *Treebanks. Building and Using Parsed Corpora*, Kluwer, Dordrecht: 189-210.

Ninio, A. 1999. Pathbreaking verbs in syntactic development and the question of prototypical transitivity. *Journal of Child Language*, 26: 619- 653.

Øvrelid L., 2004. Disambiguation of syntactic functions in Norwegian: modeling variation in word order interpretations conditioned by animacy and definiteness. *Proceedings of the 20th Scandinavian Conference of Linguistics*, Helsinki.

Quochi, V., (in preparation). A constructional analysis of parental speech: The role of frequency and prediction in language acquisition, evidence from Italian.

Ratnaparkhi A., 1998. *Maximum Entropy Models for Natural Language Ambiguity Resolution*. Ph.D. Dissertation, University of Pennsylvania.

Seidenberg M. S., MacDonald M. C. 1999. A probabilistic constraints approach to language acquisition and processing. *Cognitive Science* 23(4): 569-588.

Stemberger, J., MacWhinney, B. 1986. Frequency and the lexical storage of regularly inflected forms. *Memory and Cognition*, 14:17-26.

Tomasello M., 2000. Do young children have adult syntactic competence? *Cognition*, 74: 209-253.