Proceedings of Incontro di Grammatica Generativa XXX, Venice, Italy, 2004.

STRUCTURE AND FREQUENCY IN VERB CLASSIFICATION

Paola Merlo
University of Geneva

Suzanne Stevenson
University of Toronto

1 Introduction: the role of quantitative approaches in the formal study of language.

In this paper, we investigate the linguistic relevance of the notion of frequency in theories of lexicon organisation, in particular the definition of verb classes. Traditionally, the subject matter of linguistics has been to develop linguistic representations to describe and explain language as a cognitive process, including language acquisition and language comprehension and production. Theories that have been developed to adhere to this research plan have largely been of the symbolic, algebraic, categorical kind. Quantitative methods and corpus-based data collection have been used extensively in the study of language acquisition, language processing, historical linguistics and sociolingustics, but they have been systematically excluded from the representations and the methods used in the study of formal grammars. The only recognition of non-categorical phenomena in traditional grammars has been the notion of markedness. But that need not be. Investigations of the link between a richly structured linguistic theory and the distributional properties of language are not contradictory with the goals of generative grammar. The availability of probabilistic information has been shown to affect the learnability of a language and to enable learning with less and poorer data (Horning 1969). There is also ample evidence of frequency effects in language processing (Seidenberg, MacDonald and Saffran 2002, MacDonald, Pearlmutter and Seidenberg 1994). Therefore, an investigation of the quantitative correlates of abstract linguistic concepts might in fact enhance the goals of generative grammar by elucidating the theoretical relationship between structure and frequency.

Current development of large text repositories and syntactically annotated databases, and the exponential growth of computational and storage power, allow us to ask foundational questions on the role of frequency and quantitative data in the development of theories of grammar. Along with other researchers recently, we believe that enriching traditional structural representations with quantitative information will provide stronger data, and consequently could emit predictive hypotheses in areas that were before underspecified (Bresnan, Dingare and Manning 2001, Bod, Hay and Jannedy, 2003, Keller 2000, Manning 2003). As in other empirical sciences, linguis-

tics data can be arranged on a scale of expressiveness from: nominal – categorically discrete data that cannot be ordered on a scale, such as eye colour or subcategorisation frames; to ordinal – categorically discrete data that can be ordered on a scale, such as shades of colour; to numerically discrete – such as population size; to numerically continuous – such as body weight or probabilities. The data used currently in formal grammars is nominal. Nominal data are the least expressive as they cannot be compared or ordered, and few statistical techniques can be applied to them. Quantitative data support more elaborate theories, which take into account some non-categorical facts about language. For example, we can reach a better understanding of what the phenomenon of markedness really is. Because quantitative data is more expressive, they also support theories that have fewer a priori assumptions, without losing explanatoriness and predictiveness.

In this paper, we will show that principles of the verbal lexicon organisation – verb classes – show robust statistical regularities within and across languages, and we will hypothesize that this is because these frequencies are surface reflexes of underlying thematic regularities. If taken as a piece of data in its own right, frequency then becomes a tool for discovery of underlying abstract linguistic properties.

2 Case Study: Verb Classes

One of the most influential recent research programmes on the structure of the lexicon, Levin's (1993) work on verb classes aims at reducing the information in a lexical entry to its primitive meaning components (see also Levin 1985, Pinker 1989). Under the hypothesis that semantic properties of verbs largely determine their syntactic behaviour, the linguistic knowledge about a verb consists in its specific set of meaning components along with general relations between each meaning component and its possible syntactic expressions.

Specifically, the behaviour that Levin suggests as key to verb classification is the notion of *diathesis alternation* – an alternation in the expression of the arguments of a verb, such as, e.g., the causative/inchoative alternation in *The chef melted the butter*/*The butter melted*. Levin proposes a two-stage approach. First, the semantic classes to which verbs belong are revealed empirically by the diathesis alternations they participate in. For example, *cut* and *break* can occur in the middle alternation, while *hit* and *touch* cannot. On the other hand, *hit*, *touch* and *cut* can occur in the conative alternation, while *break* cannot. Second, once classes of verbs are individuated based on contrastive syntactic behaviour, one can propose substantive hypotheses on what meaning components best describe the observed classification. For example, verbs whose meaning requires a notion of contact can participate in

the conative alternation, while verbs that do not imply physical contact, such as *break*, cannot. Using this method, Levin classifies 3024 English verbs in approximately 200 verb classes. Work by Merlo and Stevenson (2001), like others in a computational framework, have extended this idea by showing that *statistics* over the alternants of a verb effectively capture information about its class (Lapata 1999, McCarthy 2000, Schulte im Walde 2000, Lapata and Brew 2004).

Let's look at three main verb classes of English that participate in a transitivity alternation, as indicated. In Levin's account, they are distinguished from each other because the particular transitivity alternation they occur in is different in each case; however, the allowed alternants are identical for all of them – i.e., they can all be transitive or intransitive.

Manner of Motion					% Trans Usage
(1a)	The rider	raced	the horse	past the barn	23
(1b)	The horse	raced	past the barn		77
Change of State					
(2a)	The cook	melted	the butter		40
(2b)	The butter	melted			60
Creation/Transformation					
(3a)	The contractors	built	the house		62
(3b)	The contractors	built	all summer		38

We can notice that, even though the alternations do not distinguish the verbs at the syntactic level, the alternants occur across the classes with very different frequencies. Manner of motion verbs are used transitively 23% of the time and are used intransitively 77% of the time, while change of state verbs are used transitively 40% of the time and intransitively 60%, and creation/transformation verbs are more frequently transitive (62%) and less frequently intransitive (38%). These frequencies are derived by automatic counts taken from samples of 20 verbs in each class over 65 million words of Wall Street Journal text. All the differences are statistically significant. These significantly different frequencies raise several questions about the theoretical status and the generality of these frequency facts in syntax.

Question 1: Are these frequencies linguistic facts or do frequencies vary in a way that is unrelated to the abstract linguistic description? If frequencies are linguistic data, they require explanation. In particular, we need to explain why classes participate in grammatically licensed alternations so differently.

Question 2: How general are these differences in frequency distributions? Are such differences typical of all different verb classes? Moreover, is this statistical trend predictive – i.e., is the statistical trend strong enough to be definitional of the class?

Question 3: Do these differences in frequencies hold across languages? Do they reveal some commonalities across languages?

We answer these questions in the following sections in turn. The methodology is computational and experimental. Drawing on work presented in Merlo and Stevenson 2001, we first show that frequency differentials can be systematically derived from abstract properties of the verb class. We then use automatic learning techniques to explore the amount of generality of the proposed representations and of their frequency properties, showing that frequency differentials are useful in learning several new classes and across a new language.

3 Frequency, thematic roles, and animate subjects

In this section we will introduce the notions of markedness and harmonic scales to explain the connection between different lexical semantic classes and their different frequency distributions in the use of the transitive construction.

Thematic roles and frequency Recall that the first question that we want to answer is whether these differences in the relative frequency of the transitive use across classes is related to other underlying abstract properties of the formal grammar. The answer to this question is positive. Drawing from previous work (Merlo and Stevenson 2001), we will show that the difference in frequency of transitive use is related to different thematic assignments, and eventually possibly to different underlying lexical composition processes (Hale and Keyser 1993, Stevenson and Merlo 1997).

Let's look again at the examples using the verbs in question, this time indicating the thematic assignment of the participants in the event described by the verb.

Manner of Motion

(1a) The rider raced the horse past the barn
Causal Agent *Agent*

(1b) The horse raced past the barn
Agent

Change of State

(2a) The cook melted the butter
Causal Agent *Theme*

(2b) The butter melted
Theme

Creation/Transformation

(3a) The contractors built the house
Agent *Theme*

(3b) The contractors built all summer
Agent

Manner of motion verbs are intransitive action verbs whose transitive form, as in (1a), can be the causative counterpart of the intransitive form (1b). The type of causative alternation that manner of motion verbs participate in is the "induced action alternation" according to (Levin 1993). For our thematic analysis, we note that the subject of an intransitive activity verb is specified to be an Agent. The subject of the transitive form has the label Causal Agent, which indicates that the subject role is introduced with the causing event. In a causative alternation, the semantic argument of the subject of the intransitive surfaces as the object of the transitive (Brousseau and Ritter 1991, Hale and Keyser 1993, Levin 1993, Levin and Rappaport Hovav 1995). Since for manner of motion verbs this argument has agentive properties, the alternation yields an object in the transitive form that receives an Agent role (Cruse 1972, Stevenson and Merlo 1997).

The sentences in (2) illustrate the corresponding forms of a change of state verb, *melt*. Change of state verbs are intransitive, as in (2b); the transitive counterpart for these verbs also has a causative form, as in (2a). This is the "causative/inchoative alternation" (Levin 1993). Like manner of motion verbs, the subject of a transitive change of state verb is marked as the Causal Agent. Unlike manner of motion verbs, though, the alternating argument of this class of verbs (the subject of the intransitive form that becomes the object of the transitive) is a passive entity undergoing a change of state, and is therefore a Theme.

The sentences in (3) illustrate another class of verbs that can be both transitive and intransitive, creation or transformation verbs such as *build*. These are activity verbs that exhibit a non-causative transitivity alternation, in which the object is simply optional. The thematic assignment for these verbs is simply Agent for the subject (in both transitive and intransitive forms), and Theme for the optional object. We will call these classes MOM, COS

and C/T for brevity's sake in what follows. Table 1 summarizes the difference in thematic assignments.

Table 1: Thematic assignments for classes undergoing a transitivity alternation.

Class	Transitive		Intransitive
	Subject	Object	Subject
Manner of motion (MOM)	Causal Agent	Agent	Agent
Change of state (COS)	Causal Agent	Theme	Theme
Creation/ Transformation (C/T)	Agent	Theme	Agent

Can we explain the different frequency of usage of the transitive construction for these classes, based on their properties as reflected in their thematic assignment?

Subcategorisation and frequency

The Prague school's notion of linguistic markedness (Jakobson 1939, Trubetzkoy 1939) enables us to establish a scale of markedness of these thematic assignments and make a principled prediction about their frequency of occurrence. Typical tests to determine the unmarked element of a pair or scale are: *simplicity* – the unmarked element is simpler; *distribution* – the unmarked member is more widely attested across languages; and *frequency* – the unmarked member is more frequent (Greenberg 1966, Moravcsik and Wirth 1983). The claim of markedness theory is that, once an element has been identified by one test as the unmarked element of a scale, then all other tests will be correlated. The three thematic assignments appear to be ranked on a scale by the simplicity and distribution tests, as we describe below. From this, we can conclude that frequency, as a third correlated test, is also predicted to be ranked by the same scale, and we can therefore explain the observed frequencies of the three thematic assignments.

First, transitive MOM and COS verbs have a causative meaning. Since there are two events involved in a causative form, we assume that transitivity by causation has a more complex representation than simple transitives, as in the C/T verbs. Moreover, transitive MOMs are slower to process than COS transitives (Filip Tanenhaus and Carlson 1998), and the former can cause garden path effects even when they are not ambiguous (Stevenson and Merlo

1997).[1] Transitive MOMs are therefore more complex than transitive COS verbs from a processing point of view. We have thus established a scale of complexity for these three classes in a transitive usage from most (MOM) to least (C/T) complex, with COS intermediate in complexity.

We further observe that the causative transitive of a manner of motion verb has an Agent thematic role in object position which is subordinated to the Causal Agent in subject position, yielding an unusual "double agentive" thematic structure. This lexical causativization (in contrast to analytic causativization) of manner of motion verbs, which are unergatives, is found in fewer languages than lexical causatives of change of state verbs, which are syntactically unaccusative. In asking native speakers about our verbs, we found that lexical causatives of MOM verbs are not attested in Italian, French, German, Portuguese, Gungbe (Kwa family), and Czech. On the other hand, the transitive causatives are possible for change of state verbs (i.e., where the object is a Theme) in all these languages. The typological distribution test thus indicates that transitive manner of motion verbs are a distributionally rarer phenomenon than transitive change of state verbs.

Since markedness is indicated by complexity and distributional rarity, from the above observations, we can conclude that manner of motion verbs have the most marked transitive argument structure, change of state verbs have an intermediately marked transitive argument structure, and creation/transformation verbs have the least marked transitive argument structure of the three. Under the assumptions of markedness theory outlined above, we can then account for the observed behaviour: that manner of motion verbs are the least frequent in the transitive, change of state verbs have intermediate frequency in the transitive, and creation/transformation verbs are the most frequent in the transitive.

Animacy and frequency

Are there other properties of verb classes that we can expect to surface as statistical differences? Animacy is a property for which we can expect differential statistical values typical of the class, as it reflects underlying thematic assignments. Recall the pattern of thematic assignments, in Table 1

[1] We gave a processing explanation of the fact that these verbs cause a garden path, which was grounded in a specific extension of Hale and Keyser's (1993) lexical syntax proposal. We developed a specific representation for these cases which require an extra level of embedding, hence are more complex. Combined with Stevenson's competitive processing model (Stevenson 1994), we obtained the observed effects.

above. The only non-agentive subject occurs in the intransitive form of change of state verbs, which has a Theme subject. This fact has consequences for the frequency distribution of animate subjects in these classes: we expect COS verbs to have fewer animate subjects than the other two classes because we expect that Themes are less likely to be animate. This expectation follows from a combination of recent theories on the alignment of hierarchies and the thematic and animacy properties of these classes.

Recall current proposals for the harmonic combination of hierarchies:

> Alignment: Suppose given a binary dimension D1 with a scale X>Y on its elements {X,Y}, and another dimension D2 with a scale a>b>...>z on its elements. The harmonic alignment of D1 and D2 is the pair of Harmony scales:
>
> Hx: X/a > X/b>...X/z
> Hy: Y/z>....>Y/b>Y/a
>
> (Prince and Smolensky 1993, p.136)

where ">" indicates higher harmony.

In the case of thematic roles and of animacy we have the two prominence scales:

Animacy Hierarchy 1,2>3,Proper>Human>Animate>Inanimate
(Silverstein, 1976)

Thematic hierarchy AGENT > THEME

in which the relevant values combine most harmonically as Animate/AGENT>Animate/THEME.

Consequently, according to the theory of markedness, it is less marked and therefore more frequent to express an Agent with an animate entity than to express a Theme with an animate entity. We can predict that change of state verbs (the only class with a Theme subject possibility) will therefore have a lower frequency of animate subjects than the other two classes.

The predictions concerning animacy use are fully borne out by an analysis of the data across the three classes under discussion. Table 2 shows the mean relative frequencies of the two linguistic properties we have considered: use of the transitive construction and of animate subjects.
Table 2 Mean Relative Frequencies of the Data for Two Linguistic Properties

Verb Class	Linguistic Property	
	Transitive Use	Animate Subject
Manner of Motion	23%	25%
Change of State	40%	7%
Creation/Transformation	62%	15%

The data is automatically collected (over 65 million words of text), and is therefore an approximation of actual usage. All the reported differences of mean relative frequencies are statistically significant at $p<.01$. We therefore confirm statistically all the predicted orders among the classes which were hypothesized based on the relationship of frequency of transitive use and of animacy, to underlying thematic assignments of the classes of verbs.[2]

4 Generalising to new linguistic entities: the machine learning approach to theory testing

Question 2 and question 3 in the introduction mention two ways of testing whether the observed relationship between abstract linguistic properties and frequency is an idiosyncrasy of the classes under examination or a truly general and predictive property. We ask: How well do these distributional properties generalise across new verbs, across new classes and across languages? In this section, we set up the generalisation test as an automatic classification problem. We use the ability to classify new instances as a method to test the generalising power of the correlation between defining properties of the lexical semantic classes and corresponding frequencies. We test if the statistical differences observed in the previous section are strong enough to drive an automatic learner.

Formally, we say that a computer program learns from experience E with respect to some task T and performance measure P, if its performance at task T, as measured by P, improves with E. In our case, the training experience E will be provided by a database of correctly classified verbs; the task T consists in classifying verbs unseen in E into predetermined semantic classes; and the performance measure P will be defined as the percentage of verbs correctly classified. This learning paradigm is called *supervised learning*, because of the training phase, in which the algorithm is provided examples with the correct answers. During this phase the algorithm develops rules to

[2] While we confirm all predicted orders, we also observe an unpredicted distinction between the C/T and MOM classes on the animacy feature. Investigation of the possible linguistic causes of this difference is left for future research.

describe all the training data in a compact way. A possible rule in our setting could be, for example: "*If animacy is less than 10% then verb is COS*". In the testing phase, these rules are applied to additional verb data, not included in the training phase. The accuracy of classification on the test set indicates whether the rules developed in the training phase are general enough, yielding good test accuracy, or are too specific to the training set to generalise well to other data, thus yielding bad performance in the testing phase. There are numerous algorithms for learning in a supervised setting, and many regimes for training and testing such algorithms. In the following experiments, we use a decision tree induction learning algorithm, C5.0 (Quinlan, 1993), publicly available at http://www.rulequest.com, and 10-fold cross-validation repeated 10 times as the training and testing protocol.

A *decision tree* is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a classification or *decision.* The C4.5 class of decision tree induction algorithms use information theory to decide which choices provide the best partitioning of the input training data. This algorithm has good generalisation ability on many problems and yields highly readable output in the form of symbolic rules.

Cross-validation is a training and testing protocol in which the system randomly divides the data into n parts, and then runs the learner n times, using n-1 partitions for training and the remaining one for testing. At each run of the learner, a different partition is chosen for testing. This procedure is repeated m times with a different random division of the data, and the performance measure averaged over all $n * m$ experiments. When the number of data items in each class (in our case, verbs) is relatively limited, this methodology avoids the possible bias that could result from a single random split into training and testing items.

In order to present our verbs to the algorithm, each verb is encoded as a vector in which the frequencies of the identified linguistic properties serve as statistical features, as exemplified below.

Vector template: [verb, TRANS, ANIM, class]

Example: [open, .69, .36, COS]

Results confirm that the frequency correlates of the linguistic properties illustrated in the previous section are strong enough to support learning at a very good level of performance. In a task whose random baseline is approximately 33%, as it is a three-way choice, we reach performance of 70%.

This corresponds to at least a 54% reduction in error rate over the baseline.[3] The class that is most accurately classified is the class of manner of motion verbs, indicating that its markedness is easy to spot in an automatic procedure.

An analysis of errors when the algorithm is run with access to different statistical features confirms that learning does indeed occur because of the hypothesized relation between the linguistic properties and observed frequencies, and not because of some uncontrolled artefact of the experiments. If we compare a tree in which the transitive feature is used to one in which it is not, we find that the transitive property improves the discrimination of all the classes. A tree in which animacy is not used, on the other hand, has worse identification of change of state verbs, as expected.

Thus, we can conclude that not only are the frequencies systematically related to underlying properties of a sample of observed verbs (providing descriptive statistics), but that frequency differentials are also strong enough to enable a learner to classify verbs that did not belong to the initial observed sample. These frequencies are predictive.

5 Generalising to new classes and to new languages

The classes of verbs presented in the previous section were chosen because they all undergo a transitivity alternation, and therefore their subcategorisation representation is the same. These classes, however, differ substantially and systematically in the percentage of use of the different subcategorisation frames that they license. In this section, we investigate other classes of verbs to show that they also exhibit differential frequency of use of their subcategorisation frames and the animacy of their subject, and that such differentials are strong enough to support learning in these cases as well. We show moreover that this predictive differential in frequency of use extends to subcategorisation frames that involve a prepositional phrase and is not limited to the transitive-intransitive distinction. We look at Psychological State verbs, Dative/Benefactive verbs, and Locative verbs, which we exemplify below in examples (4) to (7).

Psychological State

(4a) The rich love their money

(4b) The rich love too.

[3] This performance is achieved using a small number of other features related to the transitivity alternations, in addition to the TRANS and ANIM features we focus on here.

Dative/Benefactive

(5a) Bill sold Tom a car
(5b) Bill sold a car to Tom (dative)

(6a) Martha carved the baby a toy
(6b) Martha carved a toy for the baby (benefactive)

Locative

(7a) Jack sprayed paint on the wall
(7b) Jack sprayed the wall with paint

Table 3 Summary of Subcategorisation Frames and Thematic Assignments.

Verb Class	Alternant 1	Alternant 2
Psychological	NP V NP	NP V
	Experiencer, Stimulus	Experiencer
Dative/ Benefactive	NP V NP PP	NP V NP NP
	Agent Theme Goal/Beneficiary	Agent Goal/Beneficiary Theme
Locative	NP V NP PP	NP V NP PP
	Agent Locatum Location	Agent Location Locatum

Subcategorisation Frame Differently from the other three classes that occur in a transitive-intransitive alternation, psychological verbs describe a non-volitional state. They can occur with an understood, generic object. Dative and benefactive verbs differ from the four previous classes of verbs because one of their arguments is a prepositional phrase or an indirect object. They describe a transfer or a benefactive action. Locative verbs have the meaning of putting/removing substances or things in/from containers or on/from surfaces. The substance or thing that is moved is the locatum argument; the place (the container or surface) is the location argument. In each variant of the alternation one of the two arguments (either the locatum or the location) is expressed as the object of a preposition, while the other is expressed as a direct object.

The patterns of subcategorisation frames and thematic role assignments that distinguish these classes are shown in Table 3. One can notice that psychological verbs are simple transitives, without causation, and we predict therefore that they will be at least as frequent in the transitive form as the creation/transformation verbs. For the other two classes of verbs, we simply predict that use of particular prepositions will be a good predictor of the thematic roles assigned underlyingly.

Animacy The notion of animacy of the subject which was developed in the previous section is relevant to all the classes of verbs in question. The subjects of psychological verbs are experiencers: they are likely to be animate since they must be able to experience a psychological state. The subject of dative/benefactive verbs is volitional, since it must have the intention that the goal or beneficiary receive the possession or the benefit of the object or the action. Thus, it is preferentially animate. Locatives are activity verbs, like creation/transformation and manner of motion, and their subject is preferentially animate. Since all the new classes preferentially have animate subjects, they are predicted to be more frequently animate than change of state verbs.

The different subcategorisation and animacy properties of the classes of verbs under consideration translate into different frequency distributions from class to class over the transitivity and animacy properties, as confirmed by the counts reported in Table 4.[4]

Table 4 Different frequencies of transitive use and animacy of classes

Class	Transitive Use	Animacy of Subject
MOM	0.09	0.35
COS	0.36	0.20
C/T	0.39	0.37
PSY	0.54	0.49
D/B	0.47	0.30
LOC	0.44	0.34

Table 5 illustrates overall accuracy and class by class results, in terms of precision and recall of the verbs in a machine learning experiment using subcategorization and animacy features. Precision is a measure of accuracy of the classification, and tells us how many of the verbs that the algorithm assigns to a given class actually belong to that class. Recall is a measure of coverage of the automatic classification and tells us how many of the verbs

[4] One comment on the actual numbers in Table 4 is in order. The counts are collected automatically over a very large corpus, in this case the 100-million word British National Corpus. Counts of abstract notions such as animacy and, to a less extent, subcategorisation frame are therefore approximated. The numbers therefore are relevant only relationally and in their statistical properties, but their absolute values should not be taken to be an exact estimate of the phenomena in question. The fact that one can develop useful approximators is indeed in itself rather interesting, both from a computational point of view (Merlo and Stevenson 2001, Merlo, Stevenson, Tsang and Allaria 2002) and from the standpoint of language acquisition (Stevenson and Merlo 2001).

that actually belong to a class have been assigned to that class by the algorithm.

Table 5 Overall Performance and Class by Class Accuracy (P=precision, R=recall.)

Baseline (chance)											16.7
Best Performance using Subcategorisation and Animacy											56.7
MOM		COS		C/T		PSY		D/B		LOC	
P	R	P	R	P	R	P	R	P	R	P	R
67	40	36	80	67	40	75	60	80	80	50	40

The table shows that overall the algorithm classifies the verbs with 56.7% accuracy – that is, a 52% reduction of the error rate over the baseline. If we look at the class by class precision and recall, we observe that the D/B verbs are the best classified, because they very strongly select for the subcategorized preposition. All the other classes (except change of state verbs) have better precision than recall, in varying degrees. Change of state verbs, on the other hand, have much better recall than precision. This indicates that the algorithm has a tendency to assume that verbs are change of state, as a general rule. This is an interesting result, since the class of change of state verbs is one of the largest in Levin's classification, and does therefore constitute a very general case.

The main conclusion that we can draw from these results is that the methodology extends well to new classes, to new roles, and new subcategorisation frames. Globally, there is a reduction in error rate of 52% over the chance baseline. This indicates that the features used for the classification are of general validity, and are not limited in application to the verb classes they were initially intended for.

We present now a final set of experiments which were developed to extend the investigation to Italian, by automatically classifying Italian verbs following the same methodology as we did for English. The goal of this experiment is to verify that the observed correlation between verb classes and different frequencies are attested across languages, and that they have the same learning power that they have in English. In order to make comparisons, we set out the experiment to be as similar as possible to the previously performed experiments on English verbs. We consider five of the six classes studied for English: the psychological, change of state, creation/transformation, manner of motion, and locative verbs. We choose the particular experimental verbs by translating, as far as possible given our translation procedure, the English experimental verbs. Moreover, we use the

same features that were developed for English, to demonstrate that these properties are cross-linguistically valid, even though the Italian classes do not always allow the same alternations as their English counterparts. The training and testing regime is the same as the one described above.

The experiment is a five-way discrimination among classes that contain an equal number of verbs. Its baseline is, therefore, 20% accuracy. We obtain 50% accuracy based on the differentials of transitive use and animacy. This is a reasonably good performance, giving a 37.5% reduction in error rate.[5] Here we observe that manner of motion verbs are the best classified, while locative verbs are the worst. This result enables us to conclude that the same relationship between frequency and abstract linguistic notions related to verb classes holds for Italian, and is not, therefore, a specific property of English.

6 Conclusions

In this paper, we set out to answer three questions concerning the theoretical status of differential frequencies of some abstract syntactic properties across lexical semantic classes, their generality in a theory of lexical representation and their cross-linguistic validity. Through a set of computational experiments, we have shown that differences in frequency of transitive use and animate subjects in several classes of English and Italian verbs are systematically and predictably different.

Beside its direct relevance for a theory of lexical organisation and representation, this finding is also relevant for language acquisition studies. One of the fundamental questions of child language acquisition concerns the cues and mechanisms that are available to the child to learn the lexical semantics of the verbal lexicon. The notion of syntactic bootstrapping has been put forth, whereby the acquisition of a verb's meaning is constrained by the verb's linguistics contexts – its subcategorisation frames (Gleitman 1990) and its argument structure (Gillette, Gleitman, Gleitman and Lederer 1999). The current work is an attempt to suggest how the learner could induce subcategorisation and argument structure information. The learner uses statistics over usages that are systematically related to the underlying notion of subcategorisation frame and thematic roles, extending previous work by (Brent 1993) and (Allen 1997). We confirm the hypothesis by some very prelimi-

[5] There is reason to think that the lower absolute performance for Italian is a side-effect of the difficulty of estimating animacy automatically. Italian is a null subject language, with a clear preference for unexpressed subjects. Our estimate is based on expressed subjects and therefore is probably not as accurate as the estimate for English, as it suffers from sparse data. The development of estimates for understood elements is for future work.

nary experiments. In the context of child acquisition, we use hierarchical clustering, a more realistic method where no training phase is available (Stevenson and Merlo 2001). The frequencies we discussed above give rise to three balanced clusters distinguishing the three original classes at 63% accuracy, without supervised training. These results thus suggest that frequencies systematically correlated to underlying abstract properties of verb classes can drive lexicon acquisition.

Acknowledgments Some of the research reported in this paper was performed thanks to the generous contribution of the Swiss National Science Foundation, under grant 1114-065328 to the first author, and by the US National Science Foundation and the Natural Sciences and Engineering Research Council of Canada, under grants to the second author. We thank Eva Esteve Ferrer for her collaboration.

References

Allen, Joseph. 1997. "Probabilistic constraints in the acquisition of verb argument structure." In *Proceedings of the Third Conference of Language Acquisition*, Edinburgh, Scotland.

Bod, Rens, Jennifer Hay and Stephanie Jannedy (eds.) 2003. *Probabilistic Linguistics*, Cambridge, the MIT Press.

Brent, Michael. 1993. "From grammar to lexicon: Unsupervised learning of lexical syntax." *Computational Linguistics*, vol. 19(2), 243-262.

Bresnan, Joan, Dingare, Shipra, and Manning, Christopher 2001. "Soft constraints mirror hard constraints." In *Proceedings of the LFG01 Conference*. Stanford, CSLI Publications, 13-32.

Brousseau, Anne-Marie and Elizabeth Ritter. 1991. "A non-unified analysis of agentive verbs." In *West Coast Conference on Formal Linguistics*, 53--64.

Cruse David. 1972. "A note on English causatives." *Linguistic Inquiry*, vol.3(4), 520-528.

Filip, Hana, Michael K. Tanenhaus and Gregory N. Carlson. 1998. "Reduced Relatives Judged Hard Refute Lexical Syntactic Analysis." Talk given at the *Eleventh Annual CUNY Conference on Human Sentence Processing*. New Brunswick, N. J.

Gillette, Jane, Henry Gleitman, Lila Gleitman, and Anne Lederer, Anne. 1999. "Human simulations of vocabulary learning." *Cognition* 73, 135-176.

Gleitman, Lila. 1990. "The structural sources of verb meanings." *Language Acquisition* 1, 3-55.

Greenberg, Joseph H. 1966. *Language Universals*. The Hague, Paris, Mouton.

Gruber, Jeffrey. 1965. *Studies in Lexical Relation*. Cambridge, MA, MIT Press.

Hale, Ken and Jay Keyser. 1993. "On argument structure and the lexical representation of syntactic relations." In *The View from Building 20*, Hale and Keyser (eds.), pp. 53—110, MIT Press.

Harley, Heidi. 1995. "*Subjects, Events and Licensing*." Ph.D. thesis, Massachusetts Institute of Technology.

Horning, James 1969. *A study of grammatical inference*. PhD thesis, Stanford University.

Jakobson, Roman. 1971. "Signe Zéro." In *Selected Writings*, pp. 211-219. The Hague, Mouton, 2nd edition.

Keller, Frank 2000. *Gradience in grammar*. PhD thesis, University of Edinburgh.

Lapata, Mirella. 1999. "Acquiring lexical generalizations from corpora: A case study for diathesis alternations." In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, 397-404.

Lapata, Mirella and Chris Brew. 2004. "Verb Class Disambiguation using Informative Priors". *Computational Linguistics* 30:1, 45-73.

Levin, Beth. 1985 "Introduction." In *Lexical Semantics in Review*. Levin (ed.), pp. 1--62. Cambridge, MA, Centre for Cognitive Science, MIT.

Levin, Beth. 1993. *English Verb Classes and Alternations*. Chicago, IL, University of Chicago Press.

Levin, Beth and Malka Rappaport Hovav. 1995. *Unaccusativity*. Cambridge, MA, MIT Press.

MacDonald, Maryellen, Neal Pearlmutter, and Mark Seidenberg. (1994). "The lexical nature of syntactic ambiguity resolution." *Psychological Review, 89*, 483-506.

Manning, Christopher 2003. "Probabilistic Syntax." In Bod et al. (2003), pp.289-341. Cambridge, MA: MIT Press.

McCarthy, Diana. 2000. "Using semantic preferences to identify verbal participation in role switching alternations." In *Proceedings of ANLP-NAACL 2000*,pp. 256--263.

Merlo, Paola and Suzanne Stevenson (2001) "Automatic Verb Classification based on Statistical Distribution of Argument Structure", *Computational Linguistics*, 27:3, pp. 373--408.

Merlo, Paola, Suzanne Stevenson, Vivian Tsang and Gianluca Allaria (2002) "A Multilingual Paradigm for Automatic Verb Classification", *Procs. of the 40th Meeting of the Association for Computational Linguistics (ACL'02)*. 207-215. Philadelphia,PA.

Moravcsik, Edith and Jessica Wirth. 1983. "Markedness -- An Overview." In *Markedness*, 1-3, Fred Eckman and Edith Moravcsik and Jessica Wirth (eds.), Plenum Press, New York,NY.

Pinker, Steven. 1989. *Learnability and Cognition: the Acquisition of Argument Structure.* MIT Press, 1989.

Prince, Alan & Paul Smolensky (1993): *Optimality Theory: Constraint Interaction in Generative Grammar*. Rutgers University Center for Cognitive Science Technical Report **2**.

Quinlan, J.Ross. 1992. *C4.5 : Programs for Machine Learning*. San Mateo, CA, Morgan Kaufmann, 1992, Series in Machine Learning.

Sabine Schulte im Walde (2000). "Clustering Verbs Semantically According to their Alternation Behaviour." *Proceedings of the 18th International Conference on Computational Linguistics*, Saarbrücken, Germany, August 2000.

Seidenberg, Mark S., , Maryellen C. MacDonald and Jenny Saffran. 2002. "Does grammar start where statistics stop?" *Science*, 298, 553-554.

Silverstein,Michael, 1976. Hierarchy of features and ergativity. In Robert Dixon ed, *Grammatical Categories in Australian Languages*, Australian Institute of Aboriginal Studies, Canberra, 112-171.

Stevenson, Suzanne. 1994. "Competition and Recency in a Hybrid Network Model of Syntactic Disambiguation" *Journal of Psycholinguistic Research* 23: 4, 295-322.

Stevenson, Suzanne and Paola Merlo. 2001. "Indicator-based Learning of Argument Strucutre: Computational Experiments". Proceedings of the CUNY Human Sentence Processing Conference, Philadelphia, PA.

Stevenson, Suzanne and Paola Merlo. 1997. "Lexical structure and processing complexity." *Language and Cognitive Processes*, vol. 12, 1-2, 349--399.

Trubetzkoy, Nicolaj S. "Grundzüge der Phonologie." Prague, *Travaux du Cercle Linguistique de Prague*, 1939.