Abstraction and Generalisation in Semantic Role Labels: PropBank, VerbNet or both?

Paola Merlo
Linguistics Department
University of Geneva
5 Rue de Candolle, 1204 Geneva
Switzerland
Paola.Merlo@unige.ch

Lonneke Van Der Plas
Linguistics Department
University of Geneva
5 Rue de Candolle, 1204 Geneva
Switzerland
Lonneke.VanDerPlas@unige.ch

Abstract

Semantic role labels are the representation of the grammatically relevant aspects of a sentence meaning. Capturing the nature and the number of semantic roles in a sentence is therefore fundamental to correctly describing the interface between grammar and meaning. In this paper, we compare two annotation schemes, PropBank and VerbNet, in a task-independent, general way, analysing how well they fare in capturing the linguistic generalisations that are known to hold for semantic role labels, and consequently how well they grammaticalise aspects of meaning. We show that VerbNet is more verb-specific and better able to generalise to new semantic role instances, while PropBank better captures some of the structural constraints among roles. We conclude that these two resources should be used together, as they are complementary.

1 Introduction

Most current approaches to language analysis assume that the structure of a sentence depends on the lexical semantics of the verb and of other predicates in the sentence. It is also assumed that only certain aspects of a sentence meaning are grammaticalised. Semantic role labels are the representation of the grammatically relevant aspects of a sentence meaning.

Capturing the nature and the number of semantic roles in a sentence is therefore fundamental to correctly describe the interface between grammar and meaning, and it is of paramount importance for all natural language processing (NLP) applications that attempt to extract meaning representations from analysed text, such as question-answering systems or even machine translation.

The role of theories of semantic role lists is to obtain a set of semantic roles that can apply to any argument of any verb, to provide an unambiguous identifier of the grammatical roles of the participants in the event described by the sentence (Dowty, 1991). Starting from the first proposals (Gruber, 1965; Fillmore, 1968; Jackendoff, 1972), several approaches have been put forth, ranging from a combination of very few roles to lists of very fine-grained specificity. (See Levin and Rappaport Hovav (2005) for an exhaustive review).

In NLP, several proposals have been put forth in recent years and adopted in the annotation of large samples of text (Baker et al., 1998; Palmer et al., 2005; Kipper, 2005; Loper et al., 2007). The annotated PropBank corpus, and therefore implicitly its role labels inventory, has been largely adopted in NLP because of its exhaustiveness and because it is coupled with syntactic annotation, properties that make it very attractive for the automatic learning of these roles and their further applications to NLP tasks. However, the labelling choices made by PropBank have recently come under scrutiny (Zapirain et al., 2008; Loper et al., 2007; Yi et al., 2007).

The annotation of PropBank labels has been conceived in a two-tiered fashion. A first tier assigns abstract labels such as ARG0 or ARG1, while a separate annotation records the second-tier, verb-sense specific meaning of these labels. Labels ARG0 or ARG1 are assigned to the most prominent argument in the sentence (ARG1 for unaccusative verbs and ARG0 for all other verbs). The other labels are assigned in the order of prominence. So, while the same high-level labels are used across verbs, they could have different meanings for different verb senses. Researchers have usually concentrated on the high-level annotation, but as indicated in Yi et al. (2007), there is reason to think that these labels do not generalise across verbs, nor to unseen verbs or to novel verb

Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 288–296,
Suntec, Singapore, 2-7 August 2009.

senses. Because the meaning of the role annotation is verb-specific, there is also reason to think that it fragments the data and creates data sparseness, making automatic learning from examples more difficult. These short-comings are more apparent in the annotation of less prominent and less frequent roles, marked by the ARG2 to ARG5 labels.

Zapirain et al. (2008), Loper et al. (2007) and Yi et al. (2007) investigated the ability of the PropBank role inventory to generalise compared to the annotation in another semantic role list, proposed in the electronic dictionary VerbNet. VerbNet labels are assigned in a verb-class specific way and have been devised to be more similar to the inventories of thematic role lists usually proposed by linguists. The results in these papers are conflicting.

While Loper et al. (2007) and Yi et al. (2007) show that augmenting PropBank labels with VerbNet labels increases generalisation of the less frequent labels, such as ARG2, to new verbs and new domains, they also show that PropBank labels perform better overall, in a semantic role labelling task. Confirming this latter result, Zapirain et al. (2008) find that PropBank role labels are more robust than VerbNet labels in predicting new verb usages, unseen verbs, and they port better to new domains.

The apparent contradiction of these results can be due to several confounding factors in the experiments. First, the argument labels for which the VerbNet improvement was found are infrequent, and might therefore not have influenced the overall results enough to counterbalance new errors introduced by the finer-grained annotation scheme; second, the learning methods in both these experimental settings are largely based on syntactic information, thereby confounding learning and generalisation due to syntax — which would favour the more syntactically-driven PropBank annotation — with learning due to greater generality of the semantic role annotation; finally, task-specific learning-based experiments do not guarantee that the learners be sufficiently powerful to make use of the full generality of the semantic role labels.

In this paper, we compare the two annotation schemes, analysing how well they fare in capturing the linguistic generalisations that are known to hold for semantic role labels, and consequently how well they grammaticalise aspects of meaning. Because the well-attested strong correlation between syntactic structure and semantic role labels (Levin and Rappaport Hovav, 2005; Merlo and Stevenson, 2001) could intervene as a confounding factor in this analysis, we expressly limit our investigation to data analyses and statistical measures that do not exploit syntactic properties or parsing techniques. The conclusions reached this way are not task-specific and are therefore widely applicable.

To preview, based on results in section 3, we conclude that PropBank is easier to learn, but VerbNet is more informative in general, it generalises better to new role instances and its labels are more strongly correlated to specific verbs. In section 4, we show that VerbNet labels provide finer-grained specificity. PropBank labels are more concentrated on a few VerbNet labels at higher frequency. This is not true at low frequency, where VerbNet provides disambiguations to overloaded PropBank variables. Practically, these two sets of results indicate that both annotation schemes could be useful in different circumstances, and at different frequency bands. In section 5, we report results indicating that PropBank role sets are high-level abstractions of VerbNet role sets and that VerbNet role sets are more verb and class-specific. In section 6, we show that PropBank more closely captures the thematic hierarchy and is more correlated to grammatical functions, hence potentially more useful for semantic role labelling, for learners whose features are based on the syntactic tree. Finally, in section 7, we summarise some previous results, and we provide new statistical evidence to argue that VerbNet labels are more general across verbs. These conclusions are reached by task-independent statistical analyses. The data and the measures used to reach these conclusions are discussed in the next section.

2 Materials and Method

In data analysis and inferential statistics, careful preparation of the data and choice of the appropriate statistical measures are key. We illustrate the data and the measures used here.

2.1 Data and Semantic Role Annotation

Proposition Bank (Palmer et al., 2005) adds Levin's style predicate-argument annotation and indication of verbs' alternations to the syntactic structures of the Penn Treebank (Marcus et al.,

1993).

It defines a limited role typology. Roles are specified for each verb individually. Verbal predicates in the Penn Treebank (PTB) receive a label REL and their arguments are annotated with abstract semantic role labels A0-A5 or AA for those complements of the predicative verb that are considered arguments, while those complements of the verb labelled with a semantic functional label in the original PTB receive the composite semantic role label AM-X, where X stands for labels such as LOC, TMP or ADV, for locative, temporal and adverbial modifiers respectively. PropBank uses two levels of granularity in its annotation, at least conceptually. Arguments receiving labels A0-A5 or AA do not express consistent semantic roles and are specific to a verb, while arguments receiving an AM-X label are supposed to be adjuncts and the respective roles they express are consistent across all verbs. However, among argument labels, A0 and A1 are assigned attempting to capture Proto-Agent and Proto-Patient properties (Dowty, 1991). They are, therefore, more valid across verbs and verb instances than the A2-A5 labels. Numerical results in Yi et al. (2007) show that 85% of A0 occurrences translate into Agent roles and more than 45% instances of A1 map into Patient and Patient-like roles, using a VerbNet labelling scheme. This is also confirmed by our counts, as illustrated in Tables 3 and 4 and discussed in Section 4 below.

VerbNet is a lexical resource for English verbs, yielding argumental and thematic information (Kipper, 2005). VerbNet resembles WordNet in spirit, it provides a verbal lexicon tying verbal semantics (theta-roles and selectional restrictions) to verbal distributional syntax. VerbNet defines 23 thematic roles that are valid across verbs. The list of thematic roles can be seen in the first column of Table 4.

For some of our comparisons below to be valid, we will need to reduce the inventory of labels of VerbNet to the same number of labels in PropBank. Following previous work (Loper et al., 2007), we define equivalence classes of VerbNet labels. We will refer to these classes as VerbNet *groups*. The groups we define are illustrated in Figure 1. Notice also that all our comparisons, like previous work, will be limited to the obligatory arguments in PropBank, the A0 to A5, AA arguments, to be comparable to VerbNet. VerbNet is a lexicon and by definition it does not list optional modifiers (the arguments labelled AM-X in PropBank).

In order to support the joint use of both these resources and their comparison, SemLink has been developed (Loper et al., 2007). SemLink[1] provides mappings from PropBank to VerbNet for the WSJ portion of the Penn Treebank. The mapping have been annotated automatically by a two-stage process: a lexical mapping and an instance classifier (Loper et al., 2007). The results were hand-corrected. In addition to semantic roles for both PropBank and VerbNet, SemLink contains information about verbs, their senses and their VerbNet classes which are extensions of Levin's classes.

The annotations in SemLink 1.1. are not complete. In the analyses presented here, we have only considered occurrences of semantic roles for which both a PropBank and a VerbNet label is available in the data (roughly 45% of the PropBank semantic roles have a VerbNet semantic role).[2] Furthermore, we perform our analyses on training and development data only. This means that we left section 23 of the Wall Street Journal out. The analyses are done on the basis of 106,459 semantic role pairs.

For the analysis concerning the correlation between semantic roles and syntactic dependencies in Section 6, we merged the SemLink data with the non-projectivised gold data of the CoNNL 2008 shared task on syntactic and semantic dependency parsing (Surdeanu et al., 2008). Only those dependencies that bear both a syntactic and a semantic label have been counted for test and development set. We have discarded discontinous arguments. Analyses are based on 68,268 dependencies in total.

2.2 Measures

In the following sections, we will use simple proportions, entropy, joint entropy, conditional entropy, mutual information, and a normalised form of mutual information which measures correlation between nominal attributes called symmetric uncertainty (Witten and Frank, 2005, 291). These are all widely used measures (Manning and Schuetze, 1999), excepted perhaps the last one. We briefly describe it here.

[1] (`http://verbs.colorado.edu/semlink/`)

[2] In some cases SemLink allows for multiple annotations. In those cases we selected the first annotation.

AGENT: Agent, Agent1
PATIENT: Patient
GOAL: Recipient, Destination, Location, Source, Material, Beneficiary, Goal
EXTENT: Extent, Asset, Value
PREDATTR: Predicate, Attribute, Theme, Theme1, Theme2, Topic, Stimulus, Proposition
PRODUCT: Patient2, Product, Patient1
INSTRCAUSE: Instrument, Cause, Experiencer, Actor2, Actor, Actor1

Figure 1: VerbNet Groups

Given a random variable X, the entropy H(X) describes our uncertainty about the value of X, and hence it quantifies the information contained in a message trasmitted by this variable. Given two random variables X,Y, the joint entropy H(X,Y) describes our uncertainty about the value of the pair (X,Y). *Symmetric uncertainty* is a normalised measure of the information redundancy between the distributions of two random variables. It calculates the ratio between the joint entropy of the two random variables if they are not independent and the joint entropy if the two random variables were independent (which is the sum of their individual entropies). This measure is calculated as follows.

$$U(A,B) = 2\frac{H(A) + H(B) - H(A,B)}{H(A) + H(B)}$$

where $H(X) = -\Sigma_{x \in X}\ p(x)logp(x)$ and $H(X,Y) = -\Sigma_{x \in X, y \in Y}\ p(x,y)logp(x,y)$.

Symmetric uncertainty lies between 0 and 1. A higher value for symmetric uncertainty indicates that the two random variables are more highly associated (more redundant), while lower values indicate that the two random variables approach independence.

We use these measures to evaluate how well two semantic role inventories capture well-known distributional generalisations. We discuss several of these generalisations in the following sections.

3 Amount of Information in Semantic Roles Inventory

Most proposals of semantic role inventories agree on the fact that the number of roles should be small to be valid generally. [3]

[3] With the notable exception of FrameNet, which is developing a large number of labels organised hierarchically and interpreted frame-specifically (Ruppenhofer et al., 2006).

Task	PropBank ERR	VerbNet ERR
Role generalisation	62 (82−52/48)	66 (77−33/67)
No verbal features	48 (76−52/48)	43 (58−33/67)
Unseen predicates	50 (75−52/48)	37 (62−33/67)

Table 2: Percent Error rate reduction (ERR) across role labelling sets in three tasks in Zapirain et al. (2008). ERR= (result − baseline / 100% − baseline)

PropBank and VerbNet clearly differ in the level of granularity of the semantic roles that have been assigned to the arguments. PropBank makes fewer distinctions than VerbNet, with 7 core argument labels compared to VerbNet's 23. More important than the size of the inventory, however, is the fact that PropBank has a much more skewed distribution than VerbNet, illustrated in Table 1. Consequently, the distribution of PropBank labels has an entropy of 1.37 bits, and even when the VerbNet labels are reduced to 7 equivalence classes the distribution has an entropy of 2.06 bits. VerbNet therefore conveys more information, but it is also more difficult to learn, as it is more uncertain. An uninformed PropBank learner that simply assigned the most frequent label would be correct 52% of the times by always assigning an A1 label, while for VerbNet would be correct only 33% of the times assigning Agent.

This simple fact might cast new light on some of the comparative conclusions of previous work. In some interesting experiments, Zapirain et al. (2008) test generalising abilities of VerbNet and PropBank comparatively to new role instances in general (their Table 1, line CoNLL setting, column F1 core), and also on unknown verbs and in the absence of verbal features. They find that a learner based on VerbNet has worse learning performance. They interpret this result as indicating that VerbNet labels are less general and more dependent on knowledge of specific verbs. However, a comparison that takes into consideration the differential baseline is able to factor the difficulty of the task out of the results for the overall performance. A simple baseline for a classifier is based on a majority class assignment (see our Table 1). We use the performance results reported in Zapirain et al. (2008) and calculate the reduction in error rate based on this differential baseline for the two annotation schemes. We compare only the results for the core labels in PropBank as those

PropBank		VerbNet									
A0	38.8	Agent	32.8	Cause	1.9	Source	0.9	Asset	0.3	Goal	0.00
A1	51.7	Theme	26.3	Product	1.6	Actor1	0.8	Material	0.2	Agent1	0.00
A2	9.0	Topic	11.5	Extent	1.3	Theme2	0.8	Beneficiary	0.2		
A3	0.5	Patient	5.8	Destination	1.2	Theme1	0.8	Proposition	0.1		
A4	0.0	Experiencer	4.2	Patient1	1.2	Attribute	0.7	Value	0.1		
A5	0.0	Predicate	2.3	Location	1.0	Patient2	0.5	Instrument	0.1		
AA	0.0	Recipient	2.2	Stimulus	0.9	Actor2	0.3	Actor	0.0		

Table 1: Distribution of PropBank core labels and VerbNet labels.

are the ones that correspond to VerbNet.[4] We find more mixed results than previously reported. VerbNet has better role generalising ability overall as its reduction in error rate is greater than PropBank (first line of Table 2), but it is more degraded by lack of verb information (second and third lines of Table 2). The importance of verb information for VerbNet is confirmed by information-theoretic measures. While the entropy of VerbNet labels is higher than that of PropBank labels (2.06 bits vs. 1.37 bits), as seen before, the conditional entropy of respective PropBank and VerbNet distributions given the verb is very similar, but higher for PropBank (1.11 vs 1.03 bits), thereby indicating that the verb provides much more information in association with VerbNet labels. The mutual information of the PropBank labels and the verbs is only 0.26 bits, while it is 1.03 bits for VerbNet. These results are expected if we recall the two-tiered logic that inspired PropBank annotation, where the abstract labels are less related to verbs than labels in VerbNet.

These results lead us to our first conclusion: while PropBank is easier to learn, VerbNet is more informative in general, it generalises better to new role instances, and its labels are more strongly correlated to specific verbs. It is therefore advisable to use both annotations: VerbNet labels if the verb is available, reverting to PropBank labels if no lexical information is known.

4 Equivalence Classes of Semantic Roles

An observation that holds for all semantic role labelling schemes is that certain labels seem to be more similar than others, based on their ability to occur in the same syntactic environment and to be expressed by the same function words. For example, Agent and Instrumental Cause are often subjects (of verbs selecting animate and inanimate subjects respectively); Patients/Themes can be direct objects of transitive verbs and subjects of change of state verbs; Goal and Beneficiary can be passivised and undergo the dative alternation; Instrument and Comitative are expressed by the same preposition in many languages (see Levin and Rappaport Hovav (2005).) However, most annotation schemes in NLP and linguistics assume that semantic role labels are atomic. It is therefore hard to explain why labels do not appear to be equidistant in meaning, but rather to form equivalence classes in certain contexts. [5]

While both role inventories under scrutiny here use atomic labels, their joint distribution shows interesting relations. The proportion counts are shown in Table 3 and 4.

If we read these tables column-wise, thereby taking the more linguistically-inspired labels in VerbNet to be the reference labels, we observe that the labels in PropBank are especially concentrated on those labels that linguistically would be considered similar. Specifically, in Table 3 A0 mostly groups together Agents and Instrumental Causes; A1 mostly refers to Themes and Patients; while A2 refers to Goals and Themes. If we

[4] We assume that our majority class can roughly correspond to Zapirain et al. (2008)'s data. Notice however that both sampling methods used to collect the counts are likely to slightly overestimate frequent labels. Zapirain et al. (2008) sample only complete propositions. It is reasonable to assume that higher numbered PropBank roles (A3, A4, A5) are more difficult to define. It would therefore more often happen that these labels are not annotated than it happens that A0, A1, A2, the frequent labels, are not annotated. This reasoning is confirmed by counts on our corpus, which indicate that incomplete propositions include a higher proportion of low frequency labels and a lower proportion of high frequency labels that the overall distribution. However, our method is also likely to overestimate frequent labels, since we count all labels, even those in incomplete propositions. By the same reasoning, we will find more frequent labels than the underlying real distribution of a complete annotation.

[5] Clearly, VerbNet annotators recognise the need to express these similarities since they use variants of the same label in many cases. Because the labels are atomic however, the distance between Agent and Patient is the same as Patient and Patient1 and the intended greater similarity of certain labels is lost to a learning device. As discussed at length in the linguistic literature, features bundles instead of atomic labels would be the mechanism to capture the differential distance of labels in the inventory (Levin and Rappaport Hovav, 2005).

	A0	A1	A2	A3	A4	A5	AA
Agent	32.6	0.2	-	-	-	-	-
Patient	0.0	5.8	-	-	-	-	-
Goal	0.0	1.5	4.0	0.2	0.0	0.0	-
Extent	-	0.2	1.3	0.2	-	-	-
PredAttr	1.2	39.3	2.9	0.0	-	-	0.0
Product	0.1	2.7	0.6	-	0.0	-	-
InstrCause	4.8	2.2	0.3	0.1	-	-	-

Table 3: Distribution of PropBank by VerbNet group labels according to SemLink. Counts indicated as 0.0 approximate zero by rounding, while a - sign indicates that no occurrences were found.

read these tables row-wise, thereby concentrating on the grouping of PropBank labels provided by VerbNet labels, we see that low frequency PropBank labels are more evenly spread across VerbNet labels than the frequent labels, and it is more difficult to identify a dominant label than for high-frequency labels. Because PropBank groups together VerbNet labels at high frequency, while VerbNet labels make different distinctions at lower frequencies, the distribution of PropBank is much more skewed than VerbNet, yielding the differences in distributions and entropy discussed in the previous section.

We can draw, then, a second conclusion: while VerbNet is finer-grained than PropBank, the two classifications are not in contradiction with each other. VerbNet greater specificity can be used in different ways depending on the frequency of the label. Practically, PropBank labels could provide a strong generalisation to a VerbNet annotation at high-frequency. VerbNet labels, on the other hand, can act as disambiguators of overloaded variables in PropBank. This conclusion was also reached by Loper et al. (2007). Thus, both annotation schemes could be useful in different circumstances and at different frequency bands.

5 The Combinatorics of Semantic Roles

Semantic roles exhibit paradigmatic generalisations — generalisations across similar semantic roles in the inventory — (which we saw in section 4.) They also show syntagmatic generalisations, generalisations that concern the context. One kind of context is provided by what other roles they can occur with. It has often been observed that certain semantic roles sets are possible, while others are not; among the possible sets, certain are much more frequent than others (Levin and Rappaport Hovav, 2005). Some linguistically-inspired

	A0	A1	A2	A3	A4	A5	AA
Actor	0.0	-	-	-	-	-	-
Actor1	0.8	-	-	-	-	-	-
Actor2	-	0.3	0.1	-	-	-	-
Agent1	0.0	-	-	-	-	-	-
Agent	32.6	0.2	-	-	-	-	-
Asset	-	0.1	0.0	0.2	-	-	-
Attribute	-	0.1	0.7	-	-	-	-
Beneficiary	-	0.0	0.1	0.1	0.0	-	-
Cause	0.7	1.1	0.1	0.1	-	-	-
Destination	-	0.4	0.8	0.0	-	-	-
Experiencer	3.3	0.9	0.1	-	-	-	-
Extent	-	-	1.3	-	-	-	-
Goal	-	-	-	-	0.0	-	-
Instrument	-	-	0.1	0.0	-	-	-
Location	0.0	0.4	0.6	0.0	-	0.0	-
Material	-	0.1	0.1	0.0	-	-	-
Patient	0.0	5.8	-	-	-	-	-
Patient1	0.1	1.1	-	-	-	-	-
Patient2	-	0.1	0.5	-	-	-	-
Predicate	-	1.2	1.1	0.0	-	-	-
Product	0.0	1.5	0.1	-	0.0	-	-
Proposition	-	0.0	0.1	-	-	-	-
Recipient	-	0.3	2.0	-	0.0	-	-
Source	-	0.3	0.5	0.1	-	-	-
Stimulus	-	1.0	-	-	-	-	-
Theme	0.8	25.1	0.5	0.0	-	-	0.0
Theme1	0.4	0.4	0.0	0.0	-	-	-
Theme2	0.1	0.4	0.3	-	-	-	-
Topic	-	11.2	0.3	-	-	-	-
Value	-	0.1	-	-	-	-	-

Table 4: Distribution of PropBank by original VerbNet labels according to SemLink. Counts indicated as 0.0 approximate zero by rounding, while a - sign indicates that no occurrences were found.

semantic role labelling techniques do attempt to model these dependencies directly (Toutanova et al., 2008; Merlo and Musillo, 2008).

Both annotation schemes impose tight constraints on co-occurrence of roles, independently of any verb information, with 62 role sets for PropBank and 116 role combinations for VerbNet, fewer than possible. Among the observed role sets, some are more frequent than expected under an assumption of independence between roles. For example, in PropBank, propositions comprising A0, A1 roles are observed 85% of the time, while they would be expected to occur only in 20% of the cases. In VerbNet the difference is also great between the 62% observed Agent, PredAttr propositions and the 14% expected.

Constraints on possible role sets are the expression of structural constraints among roles inherited from syntax, which we discuss in the next section, but also of the underlying event structure of the verb. Because of this relation, we expect a strong correlation between role sets and their associated

	A0,A1	A0,A2	A1,A2
Agent, Theme	11650	109	4
Agent, Topic	8572	14	0
Agent, Patient	1873	0	0
Experiencer, Theme	1591	0	15
Agent, Product	993	1	0
Agent, Predicate	960	64	0
Experiencer, Stimulus	843	0	0
Experiencer, Cause	756	0	2

Table 5: Sample of role sets correspondences

verb, as well as role sets and verb classes for both annotation schemes. However, PropBank roles are associated based on the meaning of the verb, but also based on their positional prominence in the tree, and so we can expect their relation to the actual verb entry to be weaker.

We measure here simply the correlation as indicated by the symmetric uncertainty of the joint distribution of role sets by verbs and of role sets by verb classes, for each of the two annotation schemes. We find that the correlation between PropBank role sets and verb classes is weaker than the correlation between VerbNet role sets and verb classes, as expected (PropBank: U=0.21 vs VerbNet: U=0.46). We also find that correlation between PropBank role sets and verbs is weaker than the correlation between VerbNet role sets and verbs (PropBank: U=0.23 vs VerbNet U=0.43). Notice that this result holds for VerbNet role label groups, and is therefore not a side-effect of a different size in role inventory. This result confirms our findings reported in Table 2, which showed a larger degradation of VerbNet labels in the absence of verb information.

If we analyse the data, we see that many role sets that form one single set in PropBank are split into several sets in VerbNet, with those roles that are different being roles that in PropBank form a group. So, for example, a role list (A0, A1) in PropBank will corresponds to 14 different lists in VerbNet (when using the groups). The three most frequent VerbNet role sets describe 86% of the cases: (Agent, Predattr) 71%, (InstrCause, PredAttr) 9%, and (Agent, Patient) 6% . Using the original VerbNet labels – a very small sample of the most frequent ones is reported in Table 5 — we find 39 different sets. Conversely, we see that VerbNet sets corresponds to few PropBank sets, even for high frequency.

The third conclusion we can draw then is twofold. First, while VerbNet labels have been assigned to be valid across verbs, as confirmed by their ability to enter in many combinations, these combinations are more verb and class-specific than combinations in PropBank. Second, the fine-grained, coarse-grained correspondence of annotations between VerbNet and PropBank that was illustrated by the results in Section 4 is also borne out when we look at role sets: PropBank role sets appear to be high-level abstractions of VerbNet role sets.

6 Semantic Roles and Grammatical Functions: the Thematic Hierarchy

A different kind of context-dependence is provided by thematic hierarchies. It is a well-attested fact that lexical semantic properties described by semantic roles and grammatical functions appear to be distributed according to prominence scales (Levin and Rappaport Hovav, 2005). Semantic roles are organized according to the thematic hierarchy (one proposal among many is Agent $>$ Experiencer$>$ Goal/Source/Location$>$ Patient (Grimshaw, 1990)). This hierarchy captures the fact that the options for the structural realisation of a particular argument do not depend only on its role, but also on the roles of other arguments. For example in psychological verbs, the position of the Experiencer as a syntactic subject or object depends on whether the other role in the sentence is a Stimulus, hence lower in the hierarchy, as in the psychological verbs of the *fear* class or an Agent/Cause as in the *frighten* class. Two prominence scales can combine by matching elements harmonically, higher elements with higher elements and lower with lower (Aissen, 2003). Grammatical functions are also distributed according to a prominence scale. Thus, we find that most subjects are Agents, most objects are Patients or Themes, and most indirect objects are Goals, for example.

The semantic role inventory, thus, should show a certain correlation with the inventory of grammatical functions. However, perfect correlation is clearly not expected as in this case the two levels of representation would be linguistically and computationally redundant. Because PropBank was annotated according to argument prominence, we expect to see that PropBank reflects relationships between syntax and semantic role labels more strongly than VerbNet. Comparing syntactic dependency labels to their corresponding PropBank or VerbNet groups labels (groups are used to elim-

inate the confound of different inventory sizes), we find that the joint entropy of PropBank and dependency labels is 2.61 bits while the joint entropy of VerbNet and dependency labels is 3.32 bits. The symmetric uncertainty of PropBank and dependency labels is 0.49, while the symmetric uncertainty of VerbNet and dependency labels is 0.39.

On the basis of these correlations, we can confirm previous findings: PropBank more closely captures the thematic hierarchy and is more correlated to grammatical functions, hence potentially more useful for semantic role labelling, for learners whose features are based on the syntactic tree. VerbNet, however, provides a level of annotation that is more independent of syntactic information, a property that might be useful in several applications, such as machine translation, where syntactic information might be too language-specific.

7 Generality of Semantic Roles

Semantic roles are not meant to be domain-specific, but rather to encode aspects of our conceptualisation of the world. A semantic role inventory that wants to be linguistically perspicuous and also practically useful in several tasks needs to reflect our grammatical representation of events. VerbNet is believed to be superior in this respect to PropBank, as it attempts to be less verb-specific and to be portable across classes. Previous results (Loper et al., 2007; Zapirain et al., 2008) appear to indicate that this is not the case because a labeller has better performance with PropBank labels than with VerbNet labels. But these results are task-specific, and they were obtained in the context of parsing. Since we know that PropBank is more closely related to grammatical function and syntactic annotation than VerbNet, as indicated above in Section 6, then these results could simply indicate that parsing predicts PropBank labels better because they are more closely related to syntactic labels, and not because the semantic roles inventory is more general.

Several of the findings in the previous sections shed light on the generality of the semantic roles in the two inventories. Results in Section 3 show that previous results can be reinterpreted as indicating that VerbNet labels generalise better to new roles.

We attempt here to determine the generality of the "meaning" of a role label without recourse to a task-specific experiment. It is often claimed in the literature that semantic roles are better described by feature bundles. In particular, the features sentience and volition have been shown to be useful in distinguishing Proto-Agents from Proto-Patients (Dowty, 1991). These features can be assumed to be correlated to animacy. Animacy has indeed been shown to be a reliable indicator of semantic role differences (Merlo and Stevenson, 2001). Personal pronouns in English grammaticalise animacy. We extract all the occurrences of the unambiguously animate pronouns (*I, you, he, she, us, we, me, us, him*) and the unambiguously inanimate pronoun *it*, for each semantic role label, in PropBank and VerbNet. We find occurrences for three semantic role labels in PropBank and six in VerbNet. We reduce the VerbNet groups to five by merging Patient roles with PredAttr roles to avoid artificial variation among very similar roles. An analysis of variance of the distributions of the pronous yields a significant effect of animacy for VerbNet ($F(4)=5.62$, $p< 0.05$), but no significant effect for PropBank ($F(2)=4.94$, $p=0.11$). This result is a preliminary indication that VerbNet labels might capture basic components of meaning more clearly than PropBank labels, and that they might therefore be more general.

8 Conclusions

In this paper, we have proposed a task-independent, general method to analyse annotation schemes. The method is based on information-theoretic measures and comparison with attested linguistic generalisations, to evaluate how well semantic role inventories and annotations capture grammaticalised aspects of meaning. We show that VerbNet is more verb-specific and better able to generalise to new semantic roles, while PropBank, because of its relation to syntax, better captures some of the structural constraints among roles. Future work will investigate another basic property of semantic role labelling schemes: cross-linguistic validity.

Acknowledgements

We thank James Henderson and Ivan Titov for useful comments. The research leading to these results has received partial funding from the EU FP7 programme (FP7/2007-2013) under grant agreement number 216594 (CLASSIC project: www.classic-project.org).

References

Judith Aissen. 2003. Differential object marking: Iconicity vs. economy. *Natural Language and Linguistic Theory*, 21:435–483.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics (ACL-COLING'98)*, pages 86–90, Montreal, Canada.

David Dowty. 1991. Thematic proto-roles and argument selection. *Language*, 67(3):547–619.

Charles Fillmore. 1968. The case for case. In Emmon Bach and Harms, editors, *Universals in Linguistic Theory*, pages 1–88. Holt, Rinehart, and Winston.

Jane Grimshaw. 1990. *Argument Structure*. MIT Press.

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

Ray Jackendoff. 1972. *Semantic Interpretation in Generative Grammar*. MIT Press, Cambridge, MA.

Karin Kipper. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania.

Beth Levin and Malka Rappaport Hovav. 2005. *Argument Realization*. Cambridge University Press, Cambridge, UK.

Edward Loper, Szu ting Yi, and Martha Palmer. 2007. Combining lexical resources: Mapping between PropBank and VerbNet. In *Proceedings of the IWCS*.

Christopher Manning and Hinrich Schuetze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

Mitch Marcus, Beatrice Santorini, and M.A. Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19:313–330.

Paola Merlo and Gabriele Musillo. 2008. Semantic parsing for high-precision semantic role labelling. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning (CONLL-08)*, pages 1–8, Manchester, UK.

Paola Merlo and Suzanne Stevenson. 2001. Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics*, 27(3):373–408.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31:71–105.

Josef Ruppenhofer, Michael Ellsworth, Miriam Petruck, Christopher Johnson, and Jan Scheffczyk. 2006. Framenet ii: Theory and practice. Technical report, Berkeley,CA.

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. 2008. The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of the 12th Conference on Computational Natural Language Learning (CoNLL-2008)*, pages 159–177.

Kristina Toutanova, Aria Haghighi, and Christopher D. Manning. 2008. A global joint model for semantic role labeling. *Computational Linguistics*, 34(2).

Ian Witten and Eibe Frank. 2005. *Data Mining*. Elsevier.

Szu-ting Yi, Edward Loper, and Martha Palmer. 2007. Can semantic roles generalize across genres? In *Proceedings of the Human Language Technologies 2007 (NAACL-HLT'07)*, pages 548–555, Rochester, New York, April.

Beñat Zapirain, Eneko Agirre, and Lluís Màrquez. 2008. Robustness and generalization of role sets: PropBank vs. VerbNet. In *Proceedings of ACL-08: HLT*, pages 550–558, Columbus, Ohio, June.