Renewing and Revising SemLink

Claire Bonial, Kevin Stowe & Martha Palmer
Department of Linguistics,
University of Colorado at Boulder
Hellems 290, 295 UCB
Boulder, CO 80309-0295
{Claire.Bonial, Kevin.Stowe, Martha.Palmer}@colorado.edu

Abstract

This research describes SemLink, a comprehensive resource for Natural Language Processing that maps and unifies several high-quality lexical resources: PropBank, VerbNet, FrameNet, and the recently added OntoNotes sense groupings. Each of these resources was created for slightly different purposes, and therefore each carries unique strengths and limitations. SemLink allows users to leverage the strengths of each resource and provides the groundwork for incorporating these lexical resources effectively into linked data resources. SemLink and the resources included therein are discussed with a focus on the value of using lexical resources in a complementary fashion. Recent improvements to SemLink, including the addition of a new resource, the OntoNotes sense groupings, are described. Work to address future goals, including further expansion of SemLink, is also discussed.

1 Introduction

SemLink (Palmer, 2009) is an ongoing effort to map complementary lexical resources: PropBank (PB) (Palmer et al., 2005), VerbNet (VN) (Kipper et al., 2008), FrameNet (FN) (Fillmore et al., 2002), and the recently added OntoNotes (ON) sense groupings (Pradhan et al., 2007). Each of these lexical resources varies in the level and nature of semantic detail represented, since each was created independently with somewhat differing goals. Nonetheless, all of these resources can be used to associate semantic information with the propositions of natural language. SemLink serves as a platform to unify these resources and therefore combine the fine-granularity and rich semantics of FN, the syntactically-based generalizations of VN, and the relatively coarse-grained semantics of PB, which has been shown to be effective training data for supervised Machine Learning techniques. The recent addition of ON sense groupings, which can be thought of as a more semantically general view of WordNet (Fellbaum, 1998), provides even broader coverage for the resource.

Although SemLink has been created independently from Semantic Web technology, it is in an important tool for integrating the resources therein into linked data lexical resources, such as *lemonUby* (Eckle-Kohler, McCrae and Chiarcos, submitted). Semlink provides a single link to a lexical unit, which can then access all of these resources at once. For linked data in linguistics to be leveraged effectively, it is necessary to have systems that can automatically recognize that, for example, 'Stock prices *decreased*' and 'The stock market is *falling*' describe the same event. Such an interpretation relies upon a recognition of the similarity between *decrease* and *fall*, as well as between *stock prices* and *stock market*. This requires rich lexical resources that make these connections explicit. While WordNet and FN alone contribute much towards this goal, much more needs to be done to appropriately interpret polysemous verbs in context. SemLink helps to address this need.

SemLink unifies the aforementioned lexical resources by firstly providing a mapping between the semantic roles of PB and VN, as well as a mapping between the semantic roles of VN and the Frame Elements of FN. Each of these resources differ primarily in the granularity, or level of semantic specificity, of the semantic roles used. For example, PB uses very generic labels such as Arg0, as in:

[Arg0 President Bush] has [REL approved] [Arg1 duty-free treatment for imports of certain types of watches.]

In addition to providing several alternative syntactic frames and a set of semantic predicates corre-

sponding to verbs within a class, VN marks the PB Arg0 as an Agent, and the Arg1 as a Theme, using traditional thematic role labels. In contrast, FN labels them as Grantor and Action respectively, and puts them in the Grant Permission class, thereby situating the event within a certain semantic domain or frame. The additional semantic richness provided by VN and FN does not contradict PB, but can be seen as complementary. It should also be noted that while the explicit numbered argument label itself within PB is quite generic, PB also includes a lexical resource where these numbered arguments are further specified, and these descriptions are verb-specific and therefore quite fine-grained.

SemLink provides an additional level of unification by providing a mapping between the verb senses, or 'rolesets' of PB and VN classes, and in turn between VN classes and FN frames. Like the semantic roles, these senses also differ in their levels of granularity. For example, the verb *hear* has just one coarse-grained sense in PB, with the following roleset:

Arg0: hearer
Arg1: utterance, sound
Arg2: speaker, source of sound

This sense maps to both the Discover and See classes of VN, and the Perception_Experience and Hear frames of FN. Each resource provides a unique lexicon, again varying in the extent to which verb senses are either lumped together or distinguished. SemLink helps to leverage the contributions of each component, as well as take advantage of manual annotations created for each resource.

2 The Resources Included in SemLink

As discussed initially, the resources described here are distinct but complementary to each other. The question is, how can we best leverage the contributions of each one in a broad-coverage English lexical resource? In the quest for more annotated data and, in particular more diverse genres, it would clearly be advantageous to be able to take the manual data annotations that have been created with respect to one resource and merge them with data annotations for other resources. This could create a much larger, more diverse and yet still coherent training corpus; this is one of the goals of the SemLink project. This section provides background on each individual resource.

2.1 PropBank

Unlike FN and VN, the primary goal in developing the Proposition Bank, or PB, was not lexical resource creation, but the development of an annotated corpus to be used as training data for supervised machine learning systems. The first PB release consists of 1M words of the Wall Street Journal portion of the Penn Treebank II (Marcus & Marcinkiewicz, 1993) with predicate-argument structures for verbs, using semantic role labels for each verb argument. Although the semantic role labels are purposely chosen to be quite generic and theory neutral, Arg0, Arg1, etc., they are still intended to consistently annotate the same semantic role across syntactic variations (Arg0 and Arg1 do consistently correspond to Dowty's (1991) concepts of Proto-Agent and Proto-Patient respectively). For example, the Arg1 or Patient in 'John broke the window' is the same window that is annotated as the Arg1 in 'The window broke,' even though it is the syntactic subject in one sentence and the syntactic object in the other. Thus, the main goal of PB is to supply consistent, simple, general purpose labeling of semantic roles for a large quantity of coherent text to support the training of automatic semantic role labelers, in the same way the Penn Treebank has supported the training of statistical syntactic parsers.

As mentioned previously, PB also provides a lexicon entry for each broad meaning of every annotated verb, including the possible arguments of the predicate and their labels (its 'roleset') and all possible syntactic realizations. For example, the verb *leave* includes the following two rolesets, which correspond to syntactically and semantically distinct senses of the verb:

Roleset ID: leave.01 *move away from*
Roles:
Arg0: entity leaving
Arg1: place, person, or thing left
Arg2: attribute of arg1
Example: *John left Mary alone.*

Roleset ID: leave.02 *give*
Roles:
Arg0: giver/leaver
Arg1: thing given

Arg2: benefactive, given-to
Example: *Mary left her daughter the diamond pendant.*

This lexical resource is used as a set of verb-specific guidelines by the annotators, and can be seen as quite similar in nature to FN and VN although at a more coarse-grained level. In addition to numbered roles, PB defines several more general (ArgM, Argument Modifier) roles that can apply to any verb, and which are similar to adjuncts. These include LOCation, EXTent, ADVerbial, CAUse, TeMPoral, MaNneR, and DIRection, among others. These are marked, for example, as 'ArgM-LOC.'

In spite of its success in facilitating the training of semantic role labeling (SRL), there are several ways in which PB could be more effective. PB lacks much of the information that is contained in VN, including information about selectional restrictions, verb semantics, and inter-verb relationships. We have therefore created the mapping between VN and PB included in SemLink, which will allow us to use the machine learning techniques that have been developed for PB annotations to generate VN representations.

The mapping between VN and PB consists of two parts: a lexical mapping and an annotated corpus. The lexical mapping is responsible for specifying the potential mappings between PB and VN for a given word; but it does not specify which of those mappings (typically one to many) should be used for any given occurrence of the word. That is the job of the annotated corpus, which for any given instance gives the specific VN mapping and semantic role labels. This can be thought of as a form of sense tagging: where a PB frame maps to several VN classes, they can be thought of as more fine-grained senses, and labeling with the class label corresponds to providing a sense tag label.

The type-to-type lexical mapping was used to automatically predict VN classes and role labels for each instance. Where the resulting mapping was one-to-many, the correct mapping was selected manually (Loper et al., 2007). The usefulness of this mapping for improving SRL on new genres has been demonstrated by Yi, Loper, and Palmer (2007) who focused on Arg2 . By subdividing the Arg2 instances into coherent subgroups based on the VN labels and then using them for training, and then mapping back to Arg2 for testing, the performance on Arg2 increased 6 points for WSJ test data, and 10 points for Brown Corpus test data. These results encouraged extending the mappings to other resources, starting with FN.

2.2 VerbNet

VN is midway between PB and FN in terms of lexical specificity, and is closer to PB in its close ties to syntactic structure. It consists of hierarchically arranged verb classes, inspired by and extended from Levin's verb classes (Levin, 1993). The original Levin classes constitute the first few levels in the hierarchy, with each class subsequently refined to account for further semantic and syntactic differences within a class. In many cases, the additional information that VN provides for each class has caused it to subdivide, or use intersections of, Levin classes. Each class and subclass is characterized extensionally by its set of verbs, and intensionally by a list of the arguments of those verbs and syntactic and semantic information about them. Subclasses add information about behaviors and characteristics shared by a subset of verbs in the class.

In each class and subclass, an effort is made to list all syntactic frames in which the verbs of that class can be grammatically realized. Each syntactic frame is detailed with the expected syntactic phrase type of each argument, thematic roles of arguments, and a semantic representation; for example:

Frame NP V NP PP.destination
Example Jessica loaded boxes into the wagon.
Syntax Agent V Theme Destination
Semantics Motion(during(E), Theme)
Not(Prep-*into*(start(E), Theme, Destination))
Prep-*into*(end(E), Theme, Destination)
Cause(Agent, E)

Although this classification is primarily based on shared syntactic behaviors, there is clear semantic cohesion to each of the classes. As Levin hypothesizes, this is a result of the fact that verb behavior is a reflection of verb meaning.

2.3 FrameNet

Based on Fillmore's Frame Semantics, each semantic frame in FN is defined with respect to its Frame Elements, which are fine-grained semantic role labels. For instance, the Frame Elements for the Apply-heat Frame include a Cook, Food and

a Heating Instrument. More traditional labels for the same roles might be Agent, Theme and Instrument. Members of the Apply-heat frame include *bake, barbecue, blanch, boil, braise, broil, brown,* etc. The Apply-heat lexical units all happen to be verbs, but a frame can also have adjectives and nouns as members.

The 1,033 lexical frames are associated with over 10,000 Frame Elements, since there is a deliberate effort to keep the Frame Element names distinct whenever there are semantic differences (Fillmore et al., 2002). The Frame Elements for an individual Frame are classified in terms of how central they are, with three levels being distinguished: core (similar to syntactically obligatory), peripheral (similar to syntactically optional), and extrathematic (similar to adjuncts rather than arguments). Lexical items are grouped together based solely on having the same frame semantics, without consideration of similarity of syntactic behavior, unlike Levin's verb classes. Sets of verbs with similar syntactic behavior may appear in multiple frames, and a single FN frame may contain sets of verbs with related senses but different subcategorization properties. FN places a primary emphasis on providing rich, idiosyncratic descriptions of semantic properties of lexical units in context, and making explicit subtle differences in meaning.

The SemLink VN/FN mapping consists of three parts. The first part is a many-to-many mapping of VN Classes and FN frames for specific class members. It is many-to-many in that a given FN lexical unit can map to more than one VN member, and more frequently, a given VN member can map to more than one FN Frame. The second part is a mapping of VN semantic roles and FN frame elements. These two parts have been provided in separate files in order to offer the cleanest possible formatting. The third part is the PB corpus with mappings from PB roleset ID's to FN frames and mappings from the PB arguments to FN frame elements. This has recently been manually updated and corrected due to changes in each resource; this process is discussed in more detail in 3.1.

2.4 OntoNotes Sense Groupings

The ON Sense Groupings can be thought of as a more coarse-grained view of WordNet senses. This is because these sense groupings were based on WordNet senses that were successively merged into more coarse-grained senses based on the results of inter-annotator agreement in tagging of the senses (Duffield et al., 2007; Pradhan et al., 2007). Essentially, where two annotators were consistently able to distinguish between two senses, the distinction was kept. Where annotators were not able to consistently distinguish between two senses, the senses were conflated into one sense. For example, the sense groupings for the verb *leave* include the following 6 senses, whereas the WordNet entry includes 14 senses:

Sense 1 name='depart, go forth, exit'
Sense 2 name='leave something behind...'
Sense 3 name='cause an effect that remains'
Sense 4 name='stop, terminate, end'
Sense 5 name='exclude, neglect to include'
Sense 6 name='end a romantic relationship'

These groupings also include recently updated, manually created links to WordNet senses, VN classes and PB Framesets. Because the SemLink portion of the Wall Street Journal has also been annotated with these sense groupings, the annotation portion of SemLink has recently been augmented with the appropriate sense grouping for each instance, therefore providing an additional mapping level to the SemLink corpus. The incorporation of ON sense groupings into SemLink is discussed in more detail in 3.2.

3 Current State of SemLink

The first version of SemLink (1.1) contained mappings between the three lexical resources discussed (PB, VN, and FN), as well as a collection of predicates from the Wall Street Journal data annotated with PB and VN classes and arguments. In the recent release (SemLink 1.2, available for download here: http://verbs.colorado.edu/semlink/), these WSJ propositions have been additionally annotated with FN frames and FN frame elements (using FN version 1.5), as well as ON sense groupings. The mapping files between PB, VN (version 3.2), and FN have also been checked for consistency and updated to more accurately reflect the current relations between these resources.

3.1 FN Addition to Corpus

The first major improvement made to SemLink is the addition of FN frames and FN frame elements to the corpus annotation. SemLink 1.1 contained

mappings from VN classes to FN frames (e.g. Remove-10.1 to Change_of_leadership for class member *depose*), as well as mappings from VN thematic roles to FN frame elements (e.g. Agent to Selector for Change_of_leadership frame), but contained no FN information for specific Wall Street Journal predicates within the corpus. The current SemLink version contains manually annotated FN frames for most of these WSJ propositions, as well as automatic mappings where this was possible because the existing mapping was one-to-one. Additionally, the VN thematic role to FN frame element mapping file was used to populate the arguments for each proposition. Thus, the SemLink corpus now contains PB argument information, VN thematic roles, and the appropriately mapped FN frame elements.

The addition of FN information to the corpus data allows for a detailed inspection of these various lexical resources in language practice. The mapping files of SemLink 1.1 allowed for an overview of the granularity differences between these resources, but applying all three of them to the corpus data gives a clear picture of how each resource handles various argument structures, as well as how the resources interact and overlap with each other. With the corpus data thus annotated, a verb can be examined to see how it behaves with regard to each resource, as well as how these resources interact across a corpus.

3.2 Addition of OntoNotes Senses to SemLink

To improve and expand the variety of resources mapped by SemLink, ON sense grouping annotations were added to the corpus data in the latest SemLink release. As mentioned previously, the ON senses are derived from the WordNet sense groupings, but are more coarse-grained and allow for better inter-annotator agreement. Sense distinctions with this level of granularity can be detected automatically at 87-89% accuracy, making them effective for NLP applications (Dligach and Palmer, 2011). The coverage of ON annotations isn't complete - only 37,389 of approximately 80,000 have this annotation (although surely some of these are monosemous verbs). The current annotation covers all verbs with more than three senses and is therefore quite useful despite its incomplete coverage, but further annotation is necessary to complete the mapping of this resource.

3.3 Updates & Corrections

A pressing challenge for the SemLink project is keeping the resources that it maps properly aligned. The three major lexical resources undergo frequent revisions to improve accuracy and coverage, and the mappings between them subsequently require updates and improvements. SemLink 1.2 contains a large amount of manual updates between the mappings as well as improvements to the processes used to keep these resources aligned in the future.

The VN to FN mapping files are incredibly useful but are also challenging. Maintaining the accuracy and completeness of the files is particularly difficult, as neither resource maintains an explicit connection to the other. The mapping files between these resources were originally created and curated by hand, so that as these resources have been updated, the mapping files fall out of date. The development of SemLink 1.2 required an implementation of error checking in these files, which would indicate which VN classes, FN frames, VN thematic roles, and FN frame elements were no longer present. This allowed for these files to be checked for explicit errors and brought up to date with the current releases of both resources.

The mapping file between VN and PB contained similar errors, as both PB and VN are frequently revised, but a long-term solution for correcting these discrepancies has been developed. PB contains within its framesets explicit, hand-annotated mappings between PB frames and VN classes. The VN to PB mapping file was generated from these annotations, giving a current, accurate version of the mappings between these two resources.

With the updates to all three resources and their mapping files, the Wall Street Journal predicates were also found to contain errors resulting from antiquated annotations. Approximately one third of the instances from the original VN to PB WSJ mappings in the original SemLink contained mappings that are no longer valid, or incorrect annotations as VN and PB have been updated. The current implementation of SemLink checks each PB roleset and VN class against the current data and mapping files, and marks it for reannotation if there are any discrepancies. In this way, the WSJ data is kept consistent with the mapping files and the current versions of each resource.

4 Leveraging SemLink

Natural Language Processing applications vary widely in their of use of resources, and different applications require different levels of granularity. Research in automatic semantic role labeling has demonstrated the importance of the level of granularity of semantic roles: Yi, Loper and Palmer (2007) and Loper et al. (2007) both demonstrate that because VN labels are more generalizable across verbs than PB labels, they are easier for semantic role labeling systems to learn; however, Merlo and Van Der Plas (2009) found that the differing levels of granularity of PB and VN were both useful, and therefore suggest complementary use of both resources.

SemLink attemps to bring together both coarse and fine-grained resources and make them easily useable and interchangable. If an application requires a fine-grained resource like FN, but the available data is annotated only with a coarse-grained resource like PB, SemLink provides a bridge to make that data useable. As the coverage of SemLink expands to more data, more lexical units, and more resources, this functionality becomes more and more useful in traversing the gap between different annotations and different resource-oriented goals. Efforts to expand and improve SemLink and some of the individual resources therein are discussed in the sections to follow.

The utility of integrating resources generally, and of SemLink in particular, is also reflected in the work on UBY (Eckle-Kohler et al., 2012; Gurevychy et al., 2012), a large scale lexical semantic resource using lexical markup framework (an ISO-standard for modeling lexical resources) to uniformly represent and combine a wide range of lexical-semantic resources, like WordNet, FN and VN, but also Wiktionary and Wikipedia in both English and German. This project made use of SemLink's mappings between VN classes and FN frames to supplement its integration of resources. The UBY project brings to light the need to expand such mappings to resources between many languages, instead of being limited to English. Ideally, SemLink could in the future integrate with or expand into such a multilingual resource, for instance by linking Arabic or Hindi PropBank rolesets.

Most recently, UBY has been converted into RDF using the *lemon* lexicon model (McCrae et al., 2012; Eckle-Kohler, McCrae and Chiarcos, submitted), to create *lemonUby*. *lemon* is a lexicon model that has been specifically developed for lexical resource integration on the Semantic Web, as part of the Linguistic Linked Open Data (LLOD) initiative, which aims to develop a Linked Open Data Subcloud of Linguistics (http://linguistics.okfn.org/resources/llod/). This resource thereby provides greater interoperability between existing lexical resources and the Semantic Web, and perhaps most importantly, addresses a gap in the LLOD cloud: although there are currently many lexical resources included in the LLOD cloud, previous efforts have not included information on syntactic behaviors and semantic roles, which are crucial for lexicalizing relational knowledge. While *lemonUby* has already taken advantage of the portions of past versions of SemLink included in UBY, continued efforts to integrate the current version of SemLink will allow for other valuable lexical information from both PropBank and the ON sense groupings to become part of the LLOD cloud.

5 Future Work: Expansion of SemLink

The primary goal for future work on SemLink is to expand the resource's coverage using the following methods. Firstly, additional annotations of the existing resources can be used to provide more comprehensive mappings. Secondly, the resources themselves can be improved to have greater coverage by adding to the types of annotation included in each. Finally, the addition of PB function tags (essentially semantic role labels) to numbered arguments allows for additional mappings. Each of these improvements is discussed in more detail in the sections to follow.

5.1 Expanding Coverage with Additional Annotations

We can firstly expand SemLink's coverage by focusing on cases where the corpus would have an annotation for one or more resources, but the mappings amongst all resources are incomplete. One of the most common cases of this type is where there is more than one FN frame associated with a particular VN class, requiring manual annotation of the most appropriate frame for a particular usage in the SemLink corpus. Approximately 50,000 of these cases have recently undergone annotation and simply require adjudication before

being added to the next SemLink release. Similarly, other current annotation efforts include supplementing ON sense annotations where there are many senses associated with a given VN class.

We can also expand coverage by simply adding to the number of predicates included in an individual resource. We have started this process by examining which are the most frequent verbs in the SemLink corpus that are not included in VN. From this examination, we have discovered 20 verbs with PB annotations that are good candidates for addition to VN because they are relatively frequent in the corpus and would therefore greatly increase the full coverage of the resource: these instances make up 14,878, or 78%, of the 19,070 SemLink instances missing VN classes. These verbs include, for example, *account, be, benefit, cite, do, finance, let, market, tend, trigger,* and *violate*. Unfortunately, many of these verbs are not included in VN currently because their addition proved to be very difficult in the existing class structure: many do not readily fit into a VN class due to unique syntactic behaviors or semantic features, such as differing semantic roles. Nonetheless, 12 of these 20 verbs have already been situated in VN. Sometimes this required augmenting the existing class and subclass structure. For example, *discuss* is now found in the Chit_Chat class of VN, after some changes to the structure. In this case, the addition forced a reconsideration of the class structure, and in turn, a more rational organization for the class overall, with verbs in each of the two sibling classes fully functional in all the frames listed. The Seem class was also reorganized to more precisely capture the behavior of verbs in that class, and accommodate the extremely common verb, *be*, previously not included in VN. In other cases, entirely new classes have been added to accommodate some of these verbs. For example, the Benefit and Become classes have recently been added to VN, in order to house members such as *benefit, profit* and common copular senses of verbs like *become* and *get*.

5.2 Expanding Coverage with New Predicate Types

The second method for expanding the coverage of SemLink is to increase the number of predicate types included, which is extremely important for NLP applications. Firstly, the same event can be expressed with different parts of speech within a language; for example, *He feared the bear; His fear of bears; He is afraid of bears*. Secondly, the same event can be expressed with different parts of speech across languages, as demonstrated by the differences in the English, Hindi, and Arabic PBs. To move beyond syntactic idiosyncrasies to a deeper level of semantic representation, all of these predicate types should be included in NLP resources.

Currently, SemLink includes only verb predicates, because VN of course consists solely of verbs and PB consists largely of verbs. FN, in comparison, also includes nouns and adjectives. To address this gap, PB annotations have increasingly focused on noun and adjective predicate annotations. Guidelines for noun annotation have been developed over the past two years (guidelines available at http://verbs.colorado.edu/propbank/EPB-Annotation-Guidelines.pdf), and there are now approximately 48,000 noun annotations (although some of these simply note that the noun is not relational in the instance), and framesets for 2,549 nouns. The framesets borrow heavily from many of the frameset choices made by NomBank (Meyers et al., 2004), although the guidelines have some significant differences. Guidelines for adjective annotation are also being developed based on pilot annotations of about 5400 adjective predicates. Framesets for these adjectives are also currently being created, with 111 existing framesets. These new rolesets include mappings to FN frames and etymologically related VN classes, which will allow for future versions of SemLink to be efficiently updated.

Although separate framesets are created for each part of speech, each roleset also contains mappings to related rolesets of other parts of speech. Thus, for example, the adjective roleset absent.01 is linked to the noun roleset absence.01 and the verb roleset absent.01. Where possible, every effort is also made to ensure that the roleset itself is the same across these different parts of speech. These links allow for the creation of a unified set of framesets that represent all etymologically related realizations of the same concept across all parts of speech. This unification of PB rolesets is underway, so future versions of SemLink will be mapped to rolesets that are not tied to a particular part of speech, but rather represent a particular concept. This also facilitates the in-

tegration of PB and the Abstract Meaning Representation annotation project, the goal of which is to create a large-scale semantics bank (Banarescu et al., 2013).

5.3 Improving SemLink with PB Function Tags

Because of the differences in granularity represented by each lexical resource, there are often differences in the number of roles represented with a given predicate. PB lists roles that are found frequently with a given predicate and FN lists both 'Core' and 'Non-Core' roles separately. VN generally limits roles to those that are more 'core,' although of course this status is always debatable. As a result, there are often more roles listed in both PB and FN than in VN, and SemLink may miss links that can be made between PB and FN roles because of the gap in VN coverage. With numbered arguments alone, it can be difficult to make generalizations about PB arguments when they do not have a mapping to a VN theta role.

To address this difficulty and facilitate further mapping between FN and PB, the PB rolesets have been augmented with 'function tags' for all numbered arguments. These tags include all of PB's ArgM labels, as well as three additional tags: Proto-Agent, Proto-Patient, and Verb-Specific. These three tags are used, respectively, for Arg0, Arg1 and other arguments that simply don't have an appropriate function tag because they are quite unique to the verb in question. Each of the numbered arguments is currently being annotated with one of these function tags, allowing for users to replace the numbered args with these tags if so desired, even where a mapping to VN doesn't exist. For example, the roleset for *buy* would include the following function tags, indicated here by 'F':

Buy.01
Arg0: *Buyer*, F=Proto-Agent
Arg1: *Thing bought*, F=Proto-Patient
Arg2: *Seller*, F=Direction (used for source args)
Arg3: *Price paid*, F=Verb Specific
Arg4: *Benefactive*, F=Goal

Many of these function tags were added deterministically by using SemLink's mapping between PB arguments and VN roles. Each of the VN roles was mapped to a particular function tag; therefore, wherever there was an existing VN role mapping, this was used to supply the appropriate function tag. Manual annotations are complete for cases where there is no VN mapping.

These function tags will help to improve PB as a stand-alone corpus by allowing for the various higher-numbered arguments to be converted into more generalizable function tags. When using PB as training data, performance on Args 0 and 1 tends to be quite good because these arguments are syntactically and semantically very coherent; however, as mentioned previously, there is no consistent relationship between Args 2-5 and specific semantic roles. The function tags will facilitate useful groupings of these higher-numbered arguments. Within SemLink, the function tags can provide another level of potentially informative comparison between the more coarse-grained PB annotations and the more fine-grained roles of VN and FN, as well as overcoming gaps where a mapping to VN doesn't exist.

6 Conclusion

SemLink is a valuable tool that unifies several of the most important and comprehensive lexical resources, thereby combining the benefits of each. This unification and the mappings between resources allow for users to select the level of granularity most appropriate to their application, and to take advantage of annotations across resources. Improvements and expansions of each of the individual lexical resources included in SemLink will assist in increasing the coverage of SemLink itself, and continual updates to SemLink will ensure its quality despite ongoing changes in each of the individual lexicons and annotations included. Such improvements and expansions will allow for users to leverage the unique contributions of each of these complementary resources as each is expanded and refined. SemLink is a reminder and a reflection of the merit found in using resources in a complementary fashion: the whole, after all, can be greater than the sum of its parts. This lesson lies at the heart of linked data in linguistics, and SemLink provides a structure for greater integration of lexical resources into the Semantic Web.

Acknowledgments

We gratefully acknowledge the support of the National Science Foundation Grant NSF-IIS-1116782, A Bayesian Approach to Dynamic Lexi-

cal Resources for Flexible Language Processing, and the support of DARPA FA8750-09-C-0179 (via BBN) Machine Reading: Ontology Induction and AMR and DARPA HR0011-11-C-0145 (via LDC) BOLT, as well as the generous assistance of Silvana Hartmann of Technische Universitat Darmstadt, an UBY collaborator. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

L. Banarescu, C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and N. Schneider 2013. Abstract Meaning Representation for Sembanking. *Proceedings of the Linguistic Annotation Workshop.*

Dmitriy Dligach and Martha Palmer. 2011. Good Seed Makes a Good Crop: Accelerating Active Learning Using Language Modeling. *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.* Portland, OR.

David Dowty. 1991. Thematic Proto-roles and Argument Selection. *Language*, 67:547–619.

C.J. Duffield, J.D. Hwang, S.W. Brown, S.E. Viewig, J. Davis and M. Palmer. 2007. Criteria for the manual grouping of verb senses. *Proceedings of the Linguistic Annotation Workshop* Prague.

Judith Eckle-Kohler, Iryna Gurevychy, Silvana Hartmann, Michael Matuschek and Christian M. Meyer 2012. UBY-LMF A Uniform Model for Standardizing Heterogeneous Lexical-Semantic Resources in ISO-LMF. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)* Istanbul, Turkey.

Judith Eckle-Kohler, John Philip McCrae and Chrstian Chiarcos. submitted. *lemonUby* - a large, interlinked, syntactically-rich lexical resource for ontologies. submitted to *Semantic Web Journal*

Christiane Fellbaum (Ed.) 1998. *Wordnet: An Electronic Lexical Database*. MIT press, Cambridge.

Charles J Fillmore, Christoper R Johnson, and Miriam R L Petruck. 2002. Background to FrameNet. *International Journal of Lexicography*, 16(3):235–250.

Iryna Gurevychy, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M. Meyer, and Christian Wirth. 2012. UBY A Large-Scale Unified Lexical-Semantic Resource Based on LMF. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)* Avignon, France t.

E. Joanis, Suzanne Stevenson, and D. James. 2008. A general feature space for automatic verb classification. *Natural Language Engineering*. 14(3):337-367.

Karin Kipper, Anna Korhonen, Neville Ryant and Martha Palmer. 2008. A large-scale classification of English verbs. *Language Resources and Evaluation Journal*, 42:21–40.

Anna Korhonen and T. Briscoe. 2004. Extended lexical-semantic classification of english verbs. *Proceedings of HLT/NAACL Workshop on Computational Lexical Semanticsl* Boston, Massachusetts.

Beth Levin. 1983. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago.

Edward Loper, S. Yi, and Martha Palmer. 2007. Combining lexical resources: Mapping between PropBank and VerbNet. *Proceedings of the Seventh International Workshop on Computational Semantics (IWCS-7)* Tilburg.

Marcus M. Santorini and M.A. Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank *Computational Linguistics*, 19(2):257–285.

John Philip McCrae, G. Aguado-de Cea, P. Buitelaar, P. Cimiano, T. Declerck, A. Gomez-Perez, J. Gracia, L. Hollink, E. Montiel-Ponsoda, D. Spohr, T. Wunner. 2012. Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, 46:701–719.

Merlo, P., and Van Der Plas, L. 2009. Abstraction and Generalization in Semantic Role Labels: PropBank, VerbNet or both? Proceedings of the 47th Annual Meeting of the ACL and 4th IJCNLP of the AFrameNetLP, Suntec, pp. 288-296.

Adam Meyers, R. Reeves, C. Macleod, R. Szekeley, V. Zielinska, B. Young and R. Grisham. 2004. The NomBank Project: An Interim Report *Proceedings of the Frontiers in Annotation Workshop, held in conjunction with HLT/NAACL 2004* Boston, Mass.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics* 31(1):71–106.

Martha Palmer. 2009. Semlink: Linking PropBank, VerbNet and FrameNet. *Proceedings of the Generative Lexicon Conference* Pisa, Italy.

S. Pradhan, E.H. Hovy, M. Marcus, M. Palmer, L. Ramshaw and R. Weischedel. 2007. OntoNotes: A unified relational semantic representation. *Proceedings of the First IEEE International Conference on Semantic Computing(ICSC-07)* Irvine, CA.

Yi, S., Loper, E., and Palmer, M. 2007. Can semantic roles generalize across genres? Proceedings of the HLT/NAACL-2007, Rochester, pp. 548-555.