An Approach to *Take* Multi-Word Expressions

Claire Bonial* **Meredith Green**** **Jenette Preciado**** **Martha Palmer***
*Department of Linguistics, University of Colorado at Boulder
**Institute of Cognitive Science, University of Colorado at Boulder
{Claire.Bonial,Laura.Green,Jenette.Preciado,Martha.Palmer}@colorado.edu

Abstract

This research discusses preliminary efforts to expand the coverage of the PropBank lexicon to multi-word and idiomatic expressions, such as *take one for the team*. Given overwhelming numbers of such expressions, an efficient way for increasing coverage is needed. This research discusses an approach to adding multi-word expressions to the PropBank lexicon in an effective yet semantically rich fashion. The pilot discussed here uses double annotation of *take* multi-word expressions, where annotations provide information on the best strategy for adding the multi-word expression to the lexicon. This work represents an important step for enriching the semantic information included in the PropBank corpus, which is a valuable and comprehensive resource for the field of Natural Language Processing.

1 Introduction

The PropBank (PB) corpus provides information associating semantic roles with certain syntactic structures, thereby contributing valuable training data for Natural Language Processing (NLP) applications (Palmer et al., 2005). For example, recent research shows that using semantic role information in machine translation systems improves performance (Lo, Beloucif & Wu, 2013). Despite these successes, PB could be improved with greater coverage of multi-word expressions (MWEs). The PB lexicon (http://verbs.colorado.edu/PB/framesets-english) is comprised of senses of verb, noun and adjective relations, with a listing of their semantic roles (thus a sense is referred to as a 'roleset'). Although the lexicon encompasses nearly 12,000 rolesets, relatively few of these apply to instances of MWEs. PB has previously treated language as if it were purely compositional, and has therefore lumped the majority of MWEs in with lexical verb usages. For example, annotations of the single PB sense of *take* meaning *acquire, come to have, choose, bring with you from somewhere* include MWEs such as *take measures, take comfort* and *take advantage*, and likely others. Although PB senses typically, and this sense especially, are quite coarse-grained, valuable semantic information is lost when these distinct MWEs are lumped together with other lexical senses.

The importance of coverage for MWEs is underscored by their prevalence. Jackendoff (1997:156) estimates that the number of MWEs in a speaker's lexicon is approximately equal to the number of single words, and in WordNet 1.7 (Fellbaum, 1998), 41% of the entries were MWEs (cited in Sag et al., 2002). Furthermore, Sag (2002) estimates the vocabularies of specialized domains will continue to contribute more MWEs than simplex words. For systems like PB to continue to provide adequate training data for NLP systems, coverage must extend to MWEs. The lack of coverage in this area has already become problematic for the recently developed Abstract Meaning Representation (AMR) project (Banarescu et al., 2013), which relies upon the PB lexicon, or 'frame files' as the groundwork for its annotations. As AMR and PB have extended into more informal domains, such as online discussion forums and SMS texts, the gaps in coverage of MWEs have become more and more problematic. To address this issue, this research discusses a pilot approach to increasing the coverage of the PB lexicon to a variety of MWEs involving the verb *take*, demonstrating a methodology for efficiently augmenting the lexicon with MWEs.

2 PB Background

PB annotation was developed to provide training data for supervised machine learning classifiers. It provides semantic information, including the

Proceedings of the 10th Workshop on Multiword Expressions (MWE 2014), pages 94–98,
Gothenburg, Sweden, 26-27 April 2014.

basic "who is doing what to whom," in the form of predicate-by-predicate semantic role assignments. The annotation firstly consists of the selection of a roleset, or a coarse-grained sense of the predicate, which includes a listing of the roles, expressed as generic argument numbers, associated with that sense. Here, for example, is the roleset for Take.01, mentioned previously:

Take.01: *acquire, come to have, choose, bring*
Arg0: Taker
Arg1: Thing taken
Arg2: Taken-from, source of thing taken
Arg3: Destination

These argument numbers, along with a variety of modifier tags, such as temporal and locative, are assigned to natural language sentences drawn from a variety of corpora. The roleset and example sentences serve as a guide to annotators on how to assign argument numbers to annotation instances. The goal is to assign these simple, general-purpose labels consistently across the many possible syntactic realizations of the same event participant or semantic role.

PB has recently undertaken efforts to expand the types of predicates that are annotated. Previously, annotation efforts focused on verbs, but events generally, and even the same event, can often be expressed with a variety of different parts of speech, or with MWEs. For example,

1. He fears bears.
2. His fear of bears...
3. He is afraid of bears.
4. He has a fear of bears.

Thus, it has been necessary to expand PB annotations to provide coverage for noun, adjective and complex predicates. While this greatly enriches the semantics that PB is able to capture, it has also forced the creation of an overwhelming number of new rolesets, as generally each new predicate type receives its own set of rolesets. To alleviate this, PB has opted to begin unifying frame files through a process of 'aliasing'(Bonial et al., 2014). In this process, etymologically related concepts are aliased to each other, and aliased rolesets are unified, so that there is a single roleset representing, for example the concept of 'fear,' and this roleset is used for all syntactic instantiations of that concept.

This methodology is suited to complex predicates, such as light verb constructions (LVCs), wherein the eventive noun, carrying the bulk of the event semantics, may have an etymologically related verb that is identical in its participants or semantic roles (for a description of LVC annotation, see (Hwang et al., 2010). Thus, *have a fear* above is aliased to *fear*, as *take a bath* would be aliased to *bathe*. In this research, the possibility of extending aliasing to a variety of MWEs is explored, such that *take it easy*, as in "I'm just going to take it easy on Saturday," would be aliased to the existing lexical verb roleset for *relax*. In many cases, the semantics of MWEs are quite complex, adding shades of meaning that no lexical verb quite captures. Thus, additional strategies beyond aliasing are developed; each strategy is discussed in the following sections.

3 *Take* Pilot

For the purposes of this pilot, the *take* MWEs were gathered from WordNet's MWE and phrasal verb entries (Fellbaum, 1998), the Prague Czech-English Dependency Treebank (Hajič-2012), and Afsaneh Fazly's dissertation work (Fazly, 2007). Graduate student annotators were trained to use WordNet, Sketch Engine (Kilgarriff et al., 2004) and PB to complete double-blind annotation of these MWEs as a candidate for one of the three following strategies for increasing roleset coverage: 1) Aliasing the MWE to a lexically-similar verb or noun roleset from PB, 2) proposing the creation of groups of expressions for which one or several rolesets will be created, or 3) simply designating the MWE as an idiomatic expression. First, annotators were to try to choose a verb or noun roleset from PB that most closely resembled the syntax and semantics of the MWE. Annotators also made comments as necessary for difficult cases. The annotators were considered to have agreed if the proposed lexical verb or noun alias was the same. Strategies (2) and (3) were pursued during adjudication if the annotators were unable to agree upon an appropriate alias. Each of the possible strategies for increasing coverage is discussed in turn in the following sections.

3.1 Aliasing

Aliasing involves proposing an existing roleset from PB as a suitable roleset for future MWE annotation. LVCs were the simplest of these to alias

since the eventive or stative noun predicate (e.g.: *take a **look***) may already have an existing roleset, or there is likely an existing, etymologically related verb roleset (e.g. verb roleset Look.01). Some other MWEs were not so straightforward. For instance, *take time off* does not include an etymologically related predicate that would easily encompass the semantics of the MWE, so the annotators proposed a roleset that is not as intuitive, but captures the semantics nonetheless: the roleset for the noun *vacation*. This frame allows for an Agent to take time off, and importantly, what time is taken off from: *take time off from work, school* etc. Selecting an appropriate alias is the ideal strategy for increasing coverage, because it does not require the time and effort of manually creating a new roleset or rolesets.

Both of the instances discussed above are rather simple cases, where their coverage can be addressed efficiently through aliasing. However, many MWE instances were considerably more difficult to assign to an equivalent roleset. One such example includes *take shape*, for which the annotators decided that *shape* was an appropriate roleset. Yet, *shape* does not quite cover the unique semantics of *take shape*, which lacks the possibility of an Agent. In these cases, the MWEs may still be aliased, but they should also include an semantic constraint to convey the semantic difference, such as "-Agent" Thus, in some cases, these types of semantic constraints were used for aliases that were almost adequate, but lacked some shade of meaning conveyed by the MWE. In other cases, the semantic difference between an MWE and existing lexical verb or noun roleset was too great to be captured by the addition of such constraints, thus a new roleset or group of rolesets was created to address coverage of such MWEs, as described in the next section.

3.2 Groups of Syntactically/Lexically Similar Rolesets

In cases in which it was not possible to find a single adequate alias for an MWE, a group of rolesets representing different senses of the same MWE was created. For example, *take down* can mean *to write something down, to defeat something*, or *to deconstruct something*. Thus, a group of *take_down* rolesets were added, with each roleset reflecting one of these senses.

Similarly, some of the proposed rolesets for *take* MWEs were easily subsumed under a more coarse-grained, new frame in PB. For instance, *take one's lumps* and *take it on the chin* both more or less mean *to endure or atone for*, so combining these in a coarser-grained MWE frame is both efficient and allows for valuable distinctions in terms of semantic role labeling. Namely, the Agent choosing to atone for something, and what the entity is atoning for. However, such situations in which it's possible to create new coarse-grained MWE rolesets seem to be rare. Some MWEs initially seem similar enough to combine into a single roleset, but further exploration of usages shows that they are semantically different. *Take comfort* and *take heart in* both involve improving mood, but *take heart in* might be more closely-related to *hope* in meaning, while *take comfort in* might simply mean *to cheer up*.

3.3 Idiomatic Expression Designation

In cases in which PB annotation would be very difficult for annotators, due to polysemy or semantics that cannot be conveyed by aliasing to an existing roleset, MWEs will be listed for future annotation as Idiomatic Expressions (IE), which get special treatment. This designation indicates that the MWE is so unique that it would require its own new roleset(s) in PB, and even with these rolesets, annotators may still have difficulty determining the appropriate roleset choice or sense of the MWE. As mentioned previously, creating multiple rolesets for each expression is inefficient, especially so if the rolesets manually created will be difficult to distinguish; thus, currently such cases are simply marked with the generic IE roleset.

The MWE *take the count* is an illustrative example of this type of case. Undergraduate and graduate annotators trained in linguistics tend to have difficulty with detailed sports references in annotation instances, regardless of how much context is provided. This MWE applies to several sports scenarios: one can *take the count* in boxing or *take the (full) count* in baseball, and some usages were even found for football, where many speakers would use *run down the clock*. Annotators unfamiliar with the somewhat esoteric meanings of these phrases would undoubtedly have trouble distinguishing the rolesets and arguments of the rolesets, thus *take the count* in sports contexts (as opposed to the LVC *take the count*, meaning *to count*) will simply be designated IE.

Currently, IE instances are simply set aside from the rest of the PB corpus, so as to avoid these instances adding noise to the data. In the future, these IE expressions will need to be treated individually to determine the best way to capture their unique semantics.

4 Results & Conclusions

One way of analyzing the validity of this methodology is to examine the Inter-Annotator Agreement (IAA) on the proposed alias. After the training period (in which about 60 MWEs were investigated as a group), annotators worked on double-blind annotation of 100 additional MWEs. Of these, 17 were found to be repeats of earlier MWEs. Of the remaining 83, annotators agreed on the exact alias in 32 cases, giving a rather poor, simple IAA of about 39%. However, the standards used to calculate IAA were rigid, as only instances in which the annotators aliased the multiword expressions to exactly the same lexical verb or noun roleset were counted as an agreement. Annotators often disagreed on lexical verbs, but still chose verbs that were extraordinarily similar. Take, for example, the MWE *take back*. One annotator chose to alias this MWE to *retract* while the other annotator chose *reclaim*. It is safe to say that both of these lexical verbs are equally logical choices for *take back* and have similar semantic and syntactic qualities. In other cases, annotators had discovered different senses in their research of usages, and therefore the aliases reflect different senses of the MWE. Instances like these were marked as disagreements, resulting in a misleadingly low IAA. After discussion of disagreements, IAA for these 83 MWEs rose to 78%, leaving 18 MWEs for which the annotators were unable to agree on a strategy. Annotation proceeded with an additional 76 MWEs, and for this set annotators disagreed on only 6 MWEs. This process demonstrates that although annotators may not agree on the first alias that comes to mind, they tend to agree on similar verbs that can capture the semantics of an MWE appropriately. In a final adjudication pass, adjudicators discussed the cases of disagreement with the annotators and made a final decision on the strategy to be pursued.

In all, 159 unique MWEs were examined in double-blind annotation. Of these, 21 were discarded either because annotators felt they were not truly MWEs, and could be treated compositionally, or because they were very slight variants of other MWEs. The following table shows how many of the remaining 138 MWEs were agreed upon for aliasing (and how many of these were thought to be LVCs), how many cases led to the addition of new rolesets, how many will be labeled IE in future annotation, and how many will remain classed with the existing Take senses (note that 4 MWEs were classed as having both a potential alias for LVC usages, and requiring rolesets or another strategy for other usages; for example, *take the count* discussed above). Overall, this pilot

MWE Example	Strategy	Count
take_tumble	Alias-LVC	45
take_it_easy	Alias-nonLVC	55
take_ down	Roleset(s) Created	20
take_count	IE	4
take_home	Take.XX	18

Table 1: MWE cases addressed by each strategy.

demonstrated that the approach is promising, considering that it requires only about 20 new rolesets to be created, as opposed to over 138 (given that some MWEs have multiple senses, requiring multiple rolesets). As annotations move on to additional MWEs involving other verbs, a similar reduction in the roleset workload will be invaluable to expanding PB.

5 Future Work

The next step in this research is to complete the roleset unification, which allows the aliasing to take effect. This process is currently underway. Once this is complete, an investigation of *take* annotations using the unified rolesets will be undertaken, with special focus on whether IAA for *take* instances is improved, and whether performance of automatic Semantic Role Labeling and Word Sense Disambiguation applications trained on this data is improved. If results in these areas are promising, this research will shift to analyzing *make, get,* and *have* MWEs with this methodology.

Acknowledgments

We gratefully acknowledge the support of the National Science Foundation Grant NSF-IIS-1116782, A Bayesian Approach to Dynamic Lexical Resources for Flexible Language Processing, and funding under the BOLT and Machine

Reading programs, HR0011-11-C-0145 (BOLT) FA8750-09-C-0179 (M.R.). 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.*

Claire Bonial, Julia Bonn, Kathryn Conger, Jena D. Hwang and Martha Palmer. In preparation. PropBank: Semantics of New Predicate Types. *Proceedings of the Language Resources and Evaluation Conference - LREC-2014.* Reykjavik, Iceland.

Jan Hajič, Eva Hajičov, Jarmila Panevov, Petr Sgall, Silvie Cinkov, Eva Fučkov, Marie Mikulov, Petr Pajas, Jan Popelka, Jiř Semecký, Jana Šindlerov, Jan Štěpnek, Josef Toman, Zdeňka Urešov, Zdeněk Žabokrtský. 2012. *Prague Czech-English Dependency Treebank 2.0.* Linguistic Data Consortium, Philadelphia.

Afsaneh Fazly. 2007. *Automatic Acquisition of Lexical Knowledge about Multiword Predicates*. PhD Thesis, Department of Computer Science, University of Toronto.

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

Jena D. Hwang, Archna Bhatia, Claire Bonial, Aous Mansouri, Ashwini Vaidya, Nianwen Xue and Martha Palmer. 2010. PropBank Annotation of Multilingual Light Verb Constructions *Proceedings of the Linguistic Annotation Workshop held in conjunction with ACL-2010.* Uppsala, Sweden.

Adam Kilgarriff, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. The Sketch Engine. *Proceedings of EURALEX*. Lorient, France.

Chi-kiu Lo, Meriem Beloucif, and Dekai Wu. 2013. Improving machine translation into Chinese by tuning against Chinese MEANT. *Proceedings of 10th International Workshop on Spoken Language Translation (IWSLT 2013).* Heidelberg, Germany.

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

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. *In Proceedings of the Third International Conference on Intelligent Text processing and Computational Linguistics (CICLING 2002)* 1–15. Mexico City, Mexico