Linguistic Resources for Phrasal Verb Identification

Peter A. Machonis
Florida International University
Dept. of Modern Languages
11200 SW 8th Street
Miami, FL 33199 USA
machonis@fiu.edu

Abstract

This paper shows how a lexicon grammar dictionary of English phrasal verbs (PV) can be transformed into an electronic dictionary, in order to accurately identify PV in large corpora within the linguistic development environment, NooJ. The NooJ program is an alternative to statistical methods commonly used in NLP: all PV are listed in a dictionary and then located by means of a PV grammar in both continuous and discontinuous format. Results are then refined with a series of dictionaries, disambiguating grammars, filters, and other linguistics resources. The main advantage of such a program is that all PV can be identified in any corpus. The only drawback is that PV not listed in the dictionary (e.g., archaic forms, recent neologisms) are not identified; however, new PV can easily be added to the electronic dictionary, which is freely available to all.

1 Introduction

Although described as early as the 1700's, English phrasal verbs (PV) or verb-particle combinations, such as *figure out*, *look up*, *turn on*, etc. have long been considered a characteristic trait of the English language and are to this day one of the most difficult features of English to master for non-native speakers. PV began attracting the attention of linguists in the early 1900's with Kennedy's (1920) classic study. Many have reiterated his historical analysis, such as Konishi (1958:122), who also finds a steady growth of these combinations after Old English, a slight drop during the Age of Reason – with authors such as Dryden and Johnson who avoided such "grammatical irregularities" – followed by a new expansion in the 19th century.

A renewal of interest in PV arose in the 1970's, with the works of Bolinger (1971:xi), who associated PV with a "creativeness that surpasses anything else in our language" and Fraser (1976), who first presented detailed descriptions of PV transformations. In particular, he studied constraints on particle position. Although many PV allow movement (*figure out the answer*, *figure the answer out*), if the direct object is a pronoun, it can only appear before the particle (*figure it out*, **figure out it*). Fraser (1976:19) also showed that some PV idioms prohibit particle movement (e.g., *dance up a storm*, **dance a storm up*) whereas others permit movement (e.g., *turn back the clock* or *turn the clock back*).

More recently, Hampe (2002), in her corpus-based study of semantic redundancy in English, suggests that compositional PV can function as an "index of emotional involvement of the speaker," and today, linguists and computer scientists debate the status of compositional vs. idiomatic PV. Whereas idiomatic PV, such *break up the audience* "cause to laugh" or *burn out the teacher* "exhaust," cannot be derived from the meaning of the verb plus particle and must be clearly listed in the lexicon, compositional PV, such as *drink up the milk* or *boot up the computer*, can be derived from the meaning of the regular verb. In this case, the particle simply functions as an intensifier (e.g., *rev up the engine*), aspect marker (e.g., *lock up the car*), or an adverbial noting direction (e.g., *drive up prices*). Although these are strong arguments in favor of separating compositional from idiomatic PV, Machonis (2009) suggests a disadvantage of treating compositional PV separately from frozen ones in that simple verb entries can become enormously complex when all English particles – Fraser (1976) lists fifteen different particles – are taken into account.

Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing, pages 18–27
Santa Fe, New Mexico, USA, August 20, 2018.

PV present one of the thorniest problems for Natural Language Processing; in fact, Sag et al. (2002: 14) state multiword expressions "constitute a key problem that must be resolved in order for linguistically precise NLP to succeed." This paper shows how a lexicon grammar dictionary can be transformed into an electronic dictionary, in order to correctly identify PV, both continuous and discontinuous, in large corpora, using multiple algorithms and filters within the linguistic development environment, NooJ (Silberztein, 2016).

2 Using NooJ for Automatic PV Recognition

2.1 Previous Work

Previous studies using NooJ (Machonis, 2010; 2012; 2016), showed that the automatic recognition of PV proved to be far more complex than for other multiword expressions due to three main factors: (1) their possible discontinuous nature (e.g., ***let out*** *the dogs* ⇔ ***let*** *the dogs* ***out***), (2) their confusion with verbs followed by simple prepositions (e.g., *Do you remember what I* ***asked*** *you* ***in*** *Rome?* (verb + prepositional phrase) vs. *Did you* ***ask*** *the prince* ***in*** *when he arrived?* (PV)), and (3) genuine ambiguity only resolvable from context (e.g., *Her neighbor was* ***looking over*** *the broken fence*, which can mean either "looking above the fence" (preposition) or "examining the fence" (PV)). Even with disambiguating grammars, adverbial and adjectival expression filters, and idiom dictionaries, our previous PV studies using NooJ achieved only 88% precision with written texts and 78% precision with an oral corpus, with most of the noise coming from the particles *in* and *on*, which are fairly tricky to distinguish automatically from prepositions (e.g. ***had*** *a strange smile* ***on*** *her thin lips* vs. ***had*** *her hat and jacket* ***on***), even with the disambiguation grammars, filters, and extra dictionaries mentioned above.

2.2 Using Lexicon Grammar Tables in Tandem with NooJ

The NooJ platform is a freeware linguistic development environment that can be downloaded from http://www.nooj4nlp.net/, which allows linguists to describe several levels of linguistic phenomena and then apply formalized descriptions to any corpus of texts. Instead of relying on a part of speech tagger that obligatorily produces a certain percentage of tagging mistakes, NooJ uses a Text Annotation Structure (TAS) that holds all unsolved ambiguities. Furthermore, these annotations, as opposed to tags, can represent discontinuous linguistic units, such as PV (Silberztein, 2016).

Lexicon grammar (Gross, 1994; 1996) accentuates the reproducibility of linguistic data in the form of exhaustive tables or matrices, which contain both lexical and syntactic information. For example, each verb in a table would be discussed by a team of linguists and marked as plus (+) or minus (-) for all possible complements and relevant transformations. This descriptive approach to syntax showed the enormous complexity of language and challenged the Chomskian model (Gross, 1979).

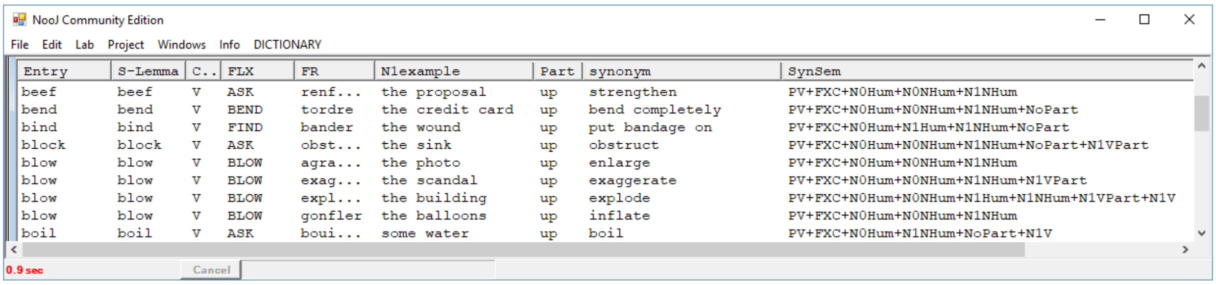
Our original PV tables included 700 entries of transitive and neutral PV[1] with the particle *up*, 200 with *out*, and 300 entries with other particles, such as *away*, *back*, *down*, *in*, *off*, *on*, *over* (300 entries). These tables are manually constructed and a sample is given in Table 1. The first two columns represent potential subjects, N_0, which can be human, non-human, or both. This is followed by the verb, the particle, and an example of a direct object, N_1. The direct object is also classified as human, non-human, or both, although only one example is given. The next column, $N_0\ V\ N_1$, takes into consideration cases where the verb can have a similar meaning, even if the particle is not used. A plus indicates that the PV can be used without the particle: e.g., *The chef beat up the eggs* ⇔ *The chef beat the eggs*. These would be considered compositional PV, since the verb keeps its regular meaning, but the particle is merely viewed as an intensifier or aspect marker, as explained in the introduction. The next column, N_1 *V Part*, identifies neutral verbs, with a plus in that column indicating that the verb has both a transitive and intransitive linked use: e.g., *She booted up the computer* ⇔ *The computer booted up*. Finally, a plus in the $N_1\ V$ column signifies that the verb can be neutral, even if the particle is not expressed: e.g., *The building blew*, *The water boiled*, *The computer booted*. The last column gives a synonym for the PV. Note that different meanings of the same PV (e.g., *beat up*, *blow up*, *bolster up*) necessitate different values in the lexicon grammar.

[1] Transitive PV take a direct object and have no intransitive: e.g., *The bully beat up the child*, but not **The bully beat up* nor **The child beat up*. Neutral PV (also referred to as ergative PV) take a direct object, which could also function as the subject: e.g., *The terrorists blew up the building* ⇔ *The building blew up*). For more information on neutral or ergative verbs within a lexicon grammar framework, see Machonis (1997).

N_0 =: Nhum	N_0 =: N-hum	Verb	Particle	Example of N_1	N_1 =: Nhum	N_1 =: N-hum	N_0 V N_1	N_1 V Part	N_1 V	Synonym
+	+	beam	up	the aliens	+	+	-	+	-	transport by energy
+	+	bear	up	the weight	+	+	+	-	-	support
+	+	beat	up	the door	-	+	-	-	-	damage
+	+	beat	up	the eggs	-	+	+	-	-	beat
+	-	beat	up	the child	+	-	+	-	-	attack physically & hurt
+	+	beef	up	the proposal	-	+	-	-	-	strengthen
+	+	bend	up	the credit card	-	+	+	-	-	bend completely
+	-	bind	up	the wound	+	+	+	-	-	put bandage on
+	+	block	up	the sink	-	+	+	+	-	obstruct
+	+	blow	up	the balloons	-	+	-	-	-	inflate
+	+	blow	up	the building	+	+	-	+	+	explode
+	+	blow	up	the photo	-	+	-	-	-	enlarge
+	+	blow	up	the scandal	-	+	-	+	-	exaggerate
+	-	boil	up	some water	-	+	+	-	+	boil
+	+	bolster	up	Max	+	+	+	-	-	give hope to
+	+	bolster	up	the theory	-	+	-	-	-	support
+	+	boot	up	the computer	-	+	+	+	+	start

Table 1: Sample Lexicon Grammar Table: Phrasal Verbs with the Particle *up*

Figure 1 below is a sample of the NooJ PV Dictionary, which mirrors all of the syntactic information contained within the highlighted area of the lexicon grammar entry in Table 1 above. As can be seen, there is also a French translation of the English PV in the NooJ dictionary. The NooJ PV Grammar in Figure 2 works in tandem with this dictionary to annotate PV in large corpora. The bottom portion of the graph represents the path for continuous PV, while the top portion of the graph represents the path for identifying discontinuous PV, i.e., with a noun phrase and optional adverb inserted between the verb and the particle. The noun phrase has an embedded NP structure, which is explained further in 3.1. Most importantly, the PV grammar uses the NooJ functionality *$THIS=$V$Part*, which assures that a particular particle must be associated with a specific verb in the PV dictionary in order for it to be recognized as a PV. That is, NooJ only recognizes verb-particle combinations listed in the PV dictionary, not simply any verb that can be part of a PV followed by any particle.



Entry	S-Lemma	C..	FLX	FR	N1example	Part	synonym	SynSem
beef	beef	V	ASK	renf...	the proposal	up	strengthen	PV+FXC+N0Hum+N0NHum+N1NHum
bend	bend	V	BEND	tordre	the credit card	up	bend completely	PV+FXC+N0Hum+N0NHum+N1NHum+NoPart
bind	bind	V	FIND	bander	the wound	up	put bandage on	PV+FXC+N0Hum+N1Hum+N1NHum+NoPart
block	block	V	ASK	obst...	the sink	up	obstruct	PV+FXC+N0Hum+N0NHum+N1NHum+NoPart+N1VPart
blow	blow	V	BLOW	agra...	the photo	up	enlarge	PV+FXC+N0Hum+N0NHum+N1NHum
blow	blow	V	BLOW	exag...	the scandal	up	exaggerate	PV+FXC+N0Hum+N0NHum+N1NHum+N1VPart
blow	blow	V	BLOW	expl...	the building	up	explode	PV+FXC+N0Hum+N0NHum+N1Hum+N1NHum+N1VPart+N1V
blow	blow	V	BLOW	gonfler	the balloons	up	inflate	PV+FXC+N0Hum+N0NHum+N1NHum
boil	boil	V	ASK	boui...	some water	up	boil	PV+FXC+N0Hum+N1NHum+NoPart+N1V

Figure 1: NooJ PV Dictionary Showing Highlighted Area of Lexicon Grammar in Table 1

In addition to the PV grammar and dictionary that work together to identify PV in large corpora, we had to add other types of resources to remove noise. These include three disambiguation grammars, which examine the immediately preceding and following environments of potential PV, and eliminate nouns that are mistaken for verbs (e.g., *take a* ***run down*** *to Spain* ≠ *run down*, *his* ***hands*** *still* ***in*** *his pockets* ≠ *hand in*), prepositions that are identified as PV (*what a comfort I* ***take in*** *it* ≠ *take in*), and prepositions that introduce locative expressions (***asked*** *you* ***in*** *Rome*). We have also written adverbial and adjectival expression filters and idiom dictionaries that identify certain fixed expressions as "unambiguous" and thus cannot be given the TAS of PV (e.g., ***asked*** ***in*** *a low tone* ≠ *ask in*, ***put on*** *one's guard* ≠ *put on*, ***take*** *an interest* ***in*** ≠ *take in*). The underlined expressions in the examples represent fixed expressions, which consequently cannot be part of a candidate PV string. The goal of these extra grammars and dictionaries is to remove noise without creating silence. More details on these resources are described in Machonis (2016).

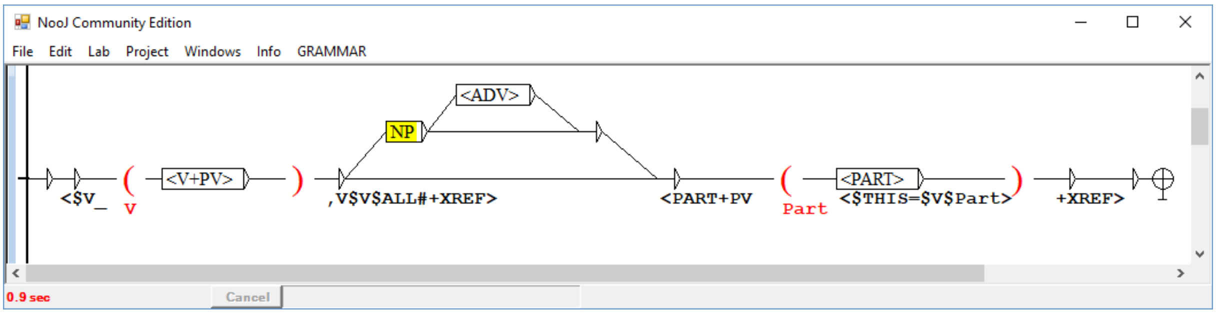


Figure 2: NooJ PV Grammar

This PV Grammar is fairly accurate, and with the recent improvements made (see section 3), can now correctly identify many discontinuous PV involving two, three, or four word forms, such as the following from our sample corpora – the 1881 Henry James novel *The Portrait of a Lady* (233,102 word forms) and an oral corpus consisting of 25 transcribed *Larry King Live* programs from January 2000 (228,950 word forms):

She had **reasoned** the matter well **out**, (*Portrait of a Lady*)
Shall I **show** the gentleman **up**, ma'am?
Mayor Ed Koch **has** a great new book **out** (*Larry King Live*)
We now know that they **tracked** it all the way **down** and then back up
That really **turned** the national economy **around**

3 Improvements to Original Grammar

To improve precision while avoiding noise, refinements were made to the original 2010 NooJ PV grammar. In this section, we present some of these enhancements.

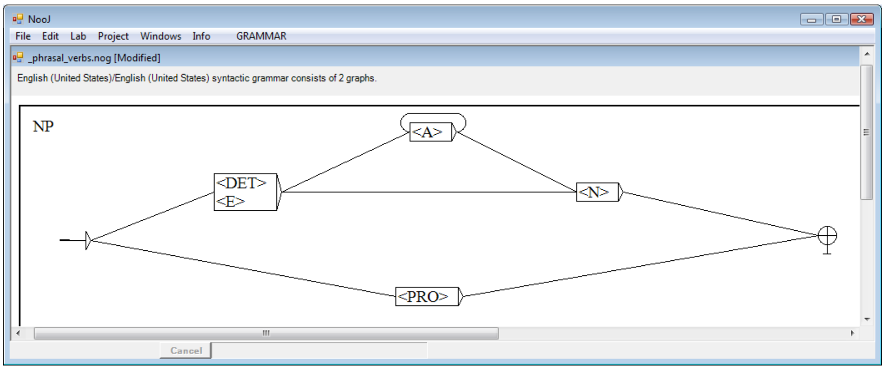
3.1 PV Grammar

One of the first things we did was to limit the embedded NP node in the PV grammar (Figure 2). While the original NP node (Figure 3) could accept a variety of noun phrases, the refined NP node (Figure 4) and DET node (Figure 5) within the present grammar restrict the type of NP allowed. For example, the sentence *he* ***had*** *found* ***out*** was previously annotated as a PV[2], i.e., *had* NP *out*, as in the example above, ***has*** *a great new book* ***out***. Now, since *found* does not have an article associated with the singular form, the verb in the sentence *he* ***had*** *found* ***out*** is no longer annotated as a PV, while the phrase ***has*** *a great new book* ***out***, where the singular NP is introduced with a determiner, still is. The new NP node is also able to identify PV originally overlooked, such as the following from our oral corpus: *and I* ***checked*** *all this* ***out***. Other PV noise removed by the new grammar includes the following from *Portrait of a Lady*:

with Pansy's little **figure** marching **up** the middle of
the band of tapestry Pansy **had** left **on** the table.
on the contrary, he **had** only let **out** sail.

[2] Although *found* is generally the past tense or past participle of the verb *find*, it can also be the present tense of the verb *found*, an adjective (*found materials*), or a noun (*they earn so much a week and found* "food and lodging in addition to pay"). NooJ recognizes all of these possibilities in the TAS.

The expression *let out*, however, is correctly identified as a PV in this last example. While disambiguation grammars and idiom dictionaries are able to remove noise automatically in NooJ, avoiding improper noun phrases is an important first step. In fact, the original 2010 PV grammar used a punctuation node after the particle to identify all discontinuous PV, which severely limited the recall of discontinuous PV. This new grammar without the punctuation node, however, had the disadvantage of identifying many prepositional phrases involving overlapping particles and prepositions as PV, which created a real difficulty for processing PV. Consequently, we added a series of disambiguation grammars.


NooJ
File Edit Lab Project Windows Info GRAMMAR
_phrasal_verbs.nog [Modified]
NP
<A>
<DET>
<E>
<N>
<PRO>
Cancel

English (United States)/English (United States) syntactic grammar consists of 2 graphs.

Figure 3: Original NP Node in NooJ PV Grammar

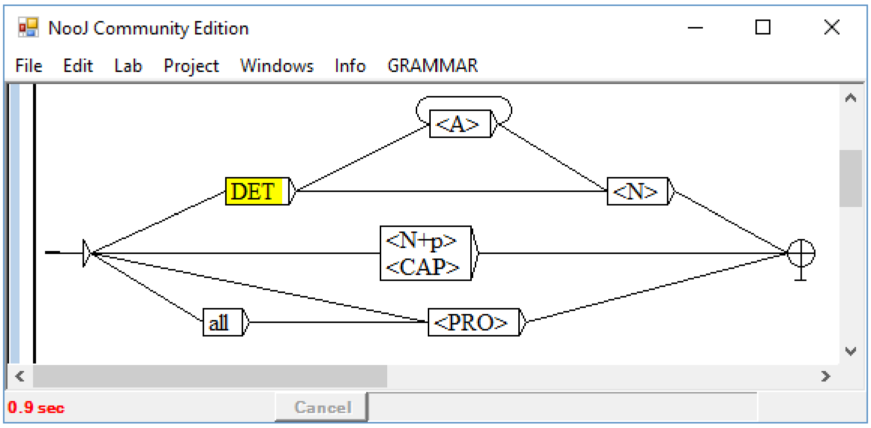


Figure 4: Revised NP Node in NooJ PV Grammar

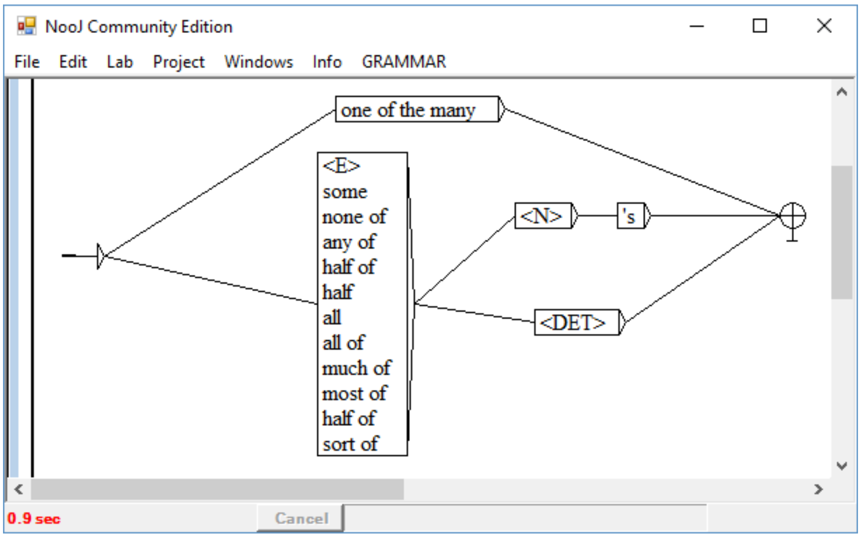


Figure 5: New DET Node in NooJ PV Grammar

3.2 Disambiguation Grammars

After a PV analysis, the TAS can be automatically modified by any of three PV disambiguation grammars, described in more detail in Machonis (2016), which specify certain structures that are **not** to be assigned PV status. The first grammar examines the environment to the right of a candidate PV string. This syntactically motivated grammar states that if the PV occurs with a pronoun object, the PV must be in the discontinuous format (e.g., *figure it out*, *look him up*, *take them away*). Thus if an object pronoun follows a supposed particle, it must be a preposition, as in the following: *what sort of pressure is* ***put on*** *them back in Cuba*. The first disambiguation grammar specifies that this instance of *put on* (e.g., *put on my T-shirt*) is not a PV. The PV *put on* is very common is our oral corpus (e.g., *put on nine pounds*, *put on my wedding dress*, *put on a prayer shawl*, *put my jeans on*), yet shows enormous potential for overlapping with prepositional phrases.

The second disambiguation grammar identifies verbs that are nouns by examining the environment to the left of a hypothetical PV. In essence, if a determiner or adjective appears immediately before the hypothetical PV, then this second disambiguation grammar correctly assumes that it is a noun and removes the PV status from the TAS. This grammar successfully eliminates much noise derived from PV that overlap with nouns, such as *break in*, *check out*, *cheer up*, *figure out*, *hand in*, *head up*, *play out*, *sort out*, *take up*, *time in*. etc.

Our third disambiguation grammar examines the environment to the right of a candidate PV string, but specifically focuses on prepositions introducing locative prepositional phrases that are clearly not part of a PV. This third disambiguation grammar makes use of a supplemental Locative Dictionary, which contains some frequent locatives found in our corpora, such as *church*, *library*, *sitting-room*, as well as place names such as *London*, *Paris*, *Rome*. These nouns are all marked as N+Loc and the PV status is automatically removed from the TAS by this grammar. For example, the place names *China* and *New Hampshire*, recently added to the Locative Dictionary, assure that the following sentences are not considered PV:

We will be **doing** it again **in** New Hampshire.
The Democrats **have** a big dispute **on** China.

Not all locative expressions have to be added to the dictionary, since some noise is already avoided by means of the new NP node in the PV grammar mentioned in 3.1 above. For example, the following represent cases of noise recently removed from our transcribed *Larry King Live* corpus. That is, the singular noun *place* does not have a determiner and the potential PV *take something in* is no longer recognized as a PV in these cases:

that trial is going to **take** place **in** Albany
process that's about to **take** place **in** Florida,
effort that will have to **take** place **in** the Pacific Ocean

3.3 Avoiding Idiom / Phrasal Verb Overlap

Although PV, as multiword expressions, are idioms in themselves, another problem we face in trying to accurately identify PV in large corpora is the overlap of certain idiomatic expressions with PV. For example, idioms that contain prepositions, such as *in*, *off*, and *on*, can easily be mistaken for PV in our corpora:

asked her **in a low tone**	≠ PV **ask in**
put the girl **on her guard**	≠ PV **put on**
take an interest in her	≠ PV **take in**

Among the additional lexical resources incorporated into NooJ, we have an Adverbials Grammar which identifies expressions such as *at one time*, *in a low tone*, *in her lap*, *on one's mind*, etc. as unambiguous adverbs (ADV+UNAMB). Consequently, neither the noun/verb (*time*), nor the preposition (*in*, *on*) can be associated with a PV.

Another grammar identifies a few idioms that also appear with certain support verbs such as *have*, *put*, and *take*, but can create noise when they are identified as PV (e.g., *have* NP *on*, *put* NP *on*, *take* NP *off*). This Adjectivals Grammar labels certain expressions as unambiguous adjectives (A+UNAMB) and consequently eliminates PV noise from sentences such as:

she **had** been much **on** her guard
the airport has already **put** grief counselors **on** duty here
was waiting to **take** him **off** his guard.

We have also incorporated into our PV analysis a larger dictionary of simple *Prep* C_1 idioms that do not have multiple modifiers, such as *in a haze*, *in the clouds*, *off duty*, *on a collision course*, *out of the question*, *over the top*, etc. These are assigned the notation A+PrepC1+UNAMB, thus avoiding noise with potential PV using the particles *in*, *off*, *on*, *out*, *over*. For example, the expression *on pins and needles* is no longer confused with the PV *keep something on* in the following sentence from *Portrait of a Lady*: *Your relations with him, while he was here, **kept** me **on** pins and needles*.

Finally, there are two dictionaries that work in tandem with grammars that target more complex idiomatic expressions such as *keep an eye on*, *take an interest in*, *take great pleasure in*, *have an opinion on*, *put blame on*, *take part in*, etc. where the frozen noun can take a variety of determiners and modifiers (Idiom1). Another dictionary lists expressions such as *turn one's back on, turn one's back against*, etc. (Idiom2). These two dictionaries work together with grammars that first identify these idioms. For example, the Idiom1 Grammar in Figure 6 annotates the expression *put the blame on* as V+Idiom1, DET, N+Idiom, PREP+Idiom. Then the Idiom Disambiguation Grammar (Figure 7) removes any potential PV from the TAS with the "this is not a PV" notation <!V+PV>, thus avoiding noise with the true PV *put something on*.

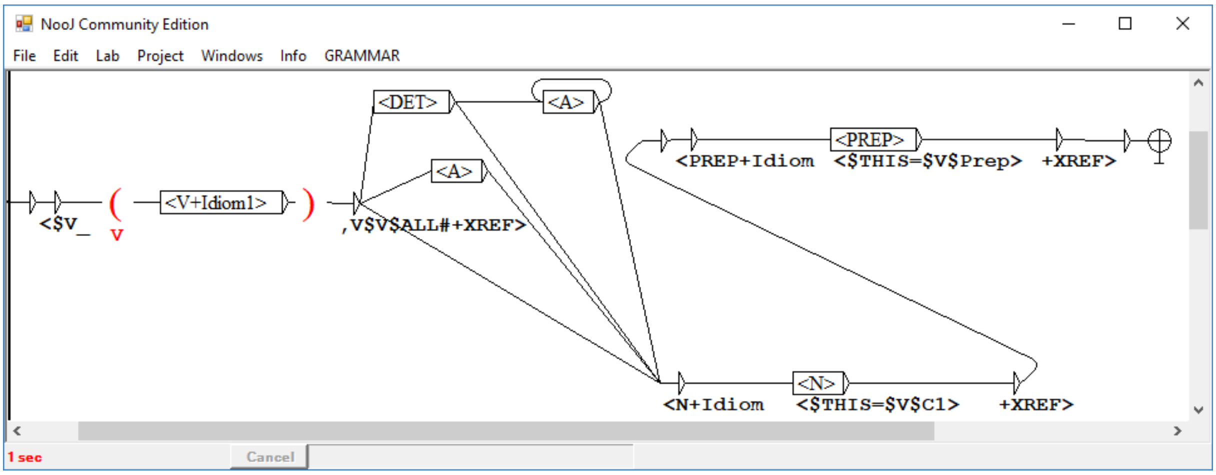


Figure 6: Verbal Idiom1 Grammar

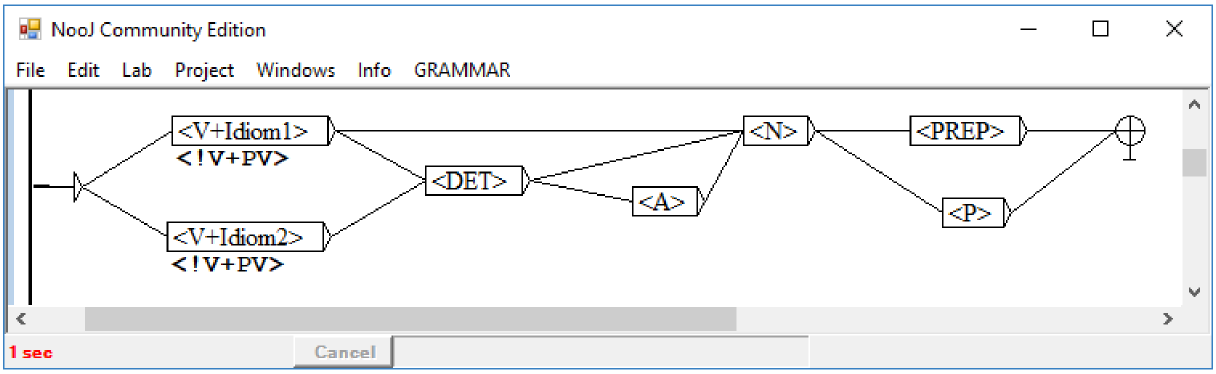


Figure 7: Idiom Disambiguation Grammar

As can be seen, most of the potential noise created by idiomatic expressions comes from the prepositions *in* and *on*, especially when used with high frequency verbs such as *keep*, *have*, *put*, *take*, and *turn*.

4 Results and Future Work

In Table 2, we present an indication of improvements made to our overall PV analysis using the two sample texts – the 1881 novel *The Portrait of a Lady* (233,102 word forms) and the 2000 oral corpus of transcribed *Larry King Live* programs (228,950 word forms). The year 2010 represents the output of our original grammar with punctuation node that overlooked many discontinuous PV. The year 2016 represents results when the punctuation node was removed from the PV Grammar and the three disambiguation grammars were added. The 2018 results represent more recent modifications to the PV Grammar, along with fine-tuning of the adverbial and adjectival expression filters and auxiliary dictionaries within NooJ. The first column represents the overall number of PV strings identified by NooJ. If a PV was automatically removed by a disambiguation grammar or other filter, it was not counted. Of the potential PV strings identified, the next two columns represent correct continuous and discontinuous PV manually verified. The next two columns represent noise, either prepositions that were incorrectly annotated as particles and part of a PV, or other PV misidentifications due to nouns mistaken for verbs (e.g., *I should spend the* ***evening out***), verbs for nouns (to **make** her reach **out** a hand), etc. The last column represents precision: True Positives / (True Positives + False Positives).

As can be seen, the number of PV automatically identified by NooJ has grown since we first started this long-term project, mainly because many discontinuous PV were not annotated due to the punctuation node requirement of the initial grammar. If this change also created much noise (false positives, incorrect PV) in our 2016 analysis, with recent changes to the PV Grammar, the addition of idiom dictionaries, and the tweaking of disambiguating grammars and other linguistics resources, precision has greatly improved, especially with our literature sample. Precision is still a major problem with our oral corpus, however, chiefly due to the noise created by the prepositions *in* and *on*. In fact, precision can be greatly improved by removing all PV with the particles *in* and *on* from the NooJ dictionary. While this does not accomplish the main NLP goal of annotating every PV in a large corpus, our resource could serve as a springboard for a purely linguistic endeavor, such as analyzing the evolution of PV throughout the history of the English language.

TEXT	Potential PV strings identified by NooJ	Correct PV (continuous)	Correct PV (discontinuous)	Incorrect PV (Prepositions)	Incorrect PV (Misidentifications)	Percentage of incorrect	PRECISION: Percentage of PV correctly identified
***Portrait of a Lady* (2010)**	583	405	83	44	51	16.30%	83.70%
***Portrait of a Lady* (2016)**	658	426	152	62	18	12.16%	87.84%
***Portrait of a Lady* (2018)**	636	426	152	55	3	9.12%	90.88%
***Larry King Live* (2010)**	614	424	102	53	35	14.33%	85.67%
***Larry King Live* (2016)**	800	451	172	136	39	21.88%	77.88%
***Larry King Live* (2018)**	730	452	169	97	12	14.93%	85.07%

Table 2: Comparison of PV Identification in 2010 & 2016 Studies vs. Today

Thim (2012:201-5), in his detailed PV study highlights “the little attention Late Modern English – in particular the 19th century – has received.” “Most of the 19th century is not covered at all,” he states. A work in progress actually involves reducing the NooJ dictionary to include only six particles (*out*, *up*, *down*, *away*, *back*, *off*) instead of twelve, which has helped us achieve 98% precision in an analysis of *The Portrait of a Lady*. In order to get a better idea of the evolution of PV in Late Modern English, we plan to use the NooJ PV Grammar, with a limited PV Dictionary, to analyze numerous 19th century texts from a variety of authors, both American (Herman Melville, James Fenimore Cooper, Washington Irving, Nathaniel Hawthorne, Harriet Beecher Stowe, Mark Twain, Edith Wharton) and British (Charles Dickens, Jane Austen, Walter Scott, the Bronte sisters, George Eliot, Thomas Hardy, Oscar Wilde).

Previously, Hiltunen (1994:135) examined English texts limiting his searches to these six typical particles representing three levels of PV frequency: high (*out*, *up*), mid (*down*, *away*), and low (*back*, *off*). By doing the same, we could get an accurate snapshot of PV usage, although limited to these six particles, in numerous 19th century novels with improved precision. In fact, our dictionary and grammar could become very useful instruments to automatically measure PV usage in different genres – novels, plays, nonfiction, technical material, daily life texts (e.g., news articles, blogs), etc. – and at different periods in the history of the English language.

In conclusion, the NooJ PV dictionary and grammar are great resources for identifying this most difficult, characteristic feature of the English language. While PV are indeed a "pain in the neck for NLP," what we have described is a reliable first step in accurately identifying them in large corpora, while automatically removing as much noise as possible. As we have seen in this paper, incorporating other idioms in an NLP analysis greatly helps to alleviate this noise. And although certain prepositions still create a fair amount of noise, many of these problems can eventually be resolved when we are able to build a NooJ grammar to recognize entire English sentences, another future goal.

References

Dwight Bolinger. 1971. *The Phrasal Verb in English*. Harvard University Press, Cambridge, MA.

Bruce Fraser. 1976. *The Verb-Particle Combination in English*. Academic Press, New York, NY.

Maurice Gross. 1979. On the failure of generative grammar. *Language* 55(4):859-885.

Maurice Gross. 1994. Constructing Lexicon grammars. In Beryl T. (Sue) Atkins & Antonio Zampolli (eds.), *Computational Approaches to the Lexicon*, 213-263. Oxford University Press, Oxford, UK.

Maurice Gross. 1996. Lexicon Grammar. In Keith Brown & Jim Miller (eds.), *Concise Encyclopedia of Syntactic Theories*, 244-258. Elsevier, New York, NY.

Beate Hampe. 2002. *Superlative Verbs: A corpus-based study of semantic redundancy in English verb-particle constructions*. Gunter Narr Verlag, Tübingen, Germany.

Risto Hiltunen. 1994. On Phrasal Verbs in Early Modern English: Notes on Lexis and Style. In Dieter Kastovsky (ed.), *Studies in Early Modern English*, 129-140. Mouton de Gruyter, Berlin, Germany.

Arthur Garfield Kennedy. 1920. *The Modern English Verb-adverb Combination*. Stanford University Press, Stanford, CA.

Tomoshichi Konishi. 1958. The growth of the verb-adverb combination in English: A brief sketch. In Kazuo Araki, Taiichiro Egawa, Toshiko Oyama & Minoru Yasui (eds.), *Studies in English grammar and linguistics: A miscellany in honour of Takanobu Otsuka,* 117-128. Kenkyusha, Tokyo, Japan.

Peter A. Machonis. 1997. Neutral verbs in English: A preliminary classification. *Lingvisticæ Investigationes* 21(2):293-320.

Peter A. Machonis. 2009. Compositional phrasal verbs with *up*: Direction, aspect, intensity. *Lingvisticae Investigationes* 32(2):253-264.

Peter A. Machonis. 2010. English Phrasal Verbs: from Lexicon Grammar to Natural Language Processing. *Southern Journal of Linguistics* 34(1):21-48.

Peter A. Machonis. 2012. *Sorting* NooJ *out* to *take* Multiword Expressions *into account.* In Kristina Vučković, Božo Bekavac, & Max Silberztein (eds.), *Automatic Processing of Various Levels of Linguistic Phenomena: Selected Papers from the NooJ 2011 International Conference*, 152-165. Cambridge Scholars Publishing, Newcastle upon Tyne, UK

Peter A. Machonis. 2016. Phrasal Verb Disambiguating Grammars: Cutting Out Noise Automatically. In Linda Barone, Max Silberztein, & Mario Monteleone (eds.), *Automatic Processing of Natural-Language Electronic Texts with NooJ*, 169-181. Springer International Publishing AG, Cham, Switzerland.

NooJ: A Linguistic Development Environment. http://www.nooj4nlp.net/

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake & Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics*, 1-15. CICLING, Mexico City, Mexico.

Max Silberztein. 2016. *Formalizing Natural Languages: The NooJ Approach.* Wiley ISTE, London, UK.

Stephan Thim. 2012. *Phrasal Verbs: The English Verb-Particle Construction and Its History*. Walter de Gruyter, Berlin, Germany.