Using Electronic Dictionaries and NooJ to Generate Sentences Containing English Phrasal Verbs

Peter A. Machonis
Florida International University
Department of Modern Languages, Miami, FL 33199, USA
machonis@fiu.edu

Abstract

This paper attempts to explore NooJ's "generation" mode to automatically produce transformations of sentences containing English Phrasal Verbs (PV). We exploit the same electronic dictionary and grammar previously used to recognize PV in large corpora (Machonis 2010, 2012), but have had to design a specific grammar for generating sentences, following the examples in Silberztein (2016), which showed how NooJ could generate over two million transformations or parallel sentences from the simple sentence *Joe likes Lea*. We created a grammar that can generate variations of a single phrase containing one of the PV found in the NooJ PV dictionary. For the moment the grammar only handles singular nouns in the present and past tense, but it is capable of applying a succession of transformations – particle movement, preterit, negation, clefting, modal insertion, aspect introduction, question formation, and passive voice, along with various combinations of these transformations – to over 1,200 PV from the electronic dictionary.

1 Introduction

English Phrasal Verbs (PV) have presented a fascinating challenge for Natural Language Processing, and as we will see in this paper, for Automatic Natural Language Generation, as well. We used the NooJ platform, a freeware linguistic development environment that can be downloaded from http://www.nooj4nlp.net/. NooJ allows linguists to describe several levels of linguistic phenomena and then apply formalized descriptions to any corpus of texts. Previously, we have used NooJ to identify all PV in large corpora, such as the complete novels of Dickens and Melville, other 19th novels, as well as a transcribed oral corpus of *Larry King Live* programs from January 2000.

Hodapp (2010), however, is the only researcher who has used NooJ to recognize PV and then apply the results for language generation. Using the NooJ PV grammar and dictionary, she designed a graphical user interface to help undergraduate students reduce PV usage – often considered informal – in academic papers. Her program generated single-word verb suggestions that could take the place of automatically identified PV.

This paper attempts to explore NooJ's "generation" mode to automatically produce paraphrases of sentences that are described by grammars. We exploit the same electronic dictionary used in NooJ to recognize PV, but have had to design a specific grammar for generating sentences that involve PV. As an initial experiment, we created a grammar that can generate variations of a single phrase containing a PV, involving transformations such as particle movement, preterit, negation, clefting – both of the subject and the object – modal insertion, aspect introduction, question formation, and passive voice, along with various combinations of these transformations. For example, from one simple sentence, such as *Max figures out the problem*, NooJ can generate over 2,500 variations such as *Didn't Max figure out the problem?*, *It was Max who started to figure the problem out*, *He should figure it out*, etc.

2 NooJ's PV Parsing Capabilities

Using NooJ, Machonis (2010, 2012) showed that the automatic recognition of PV proved to be far more complex than for other multi-word 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?* (preposition) vs. *Did you* ***ask***

Proceedings of the Linguistic Resources for Automatic Natural Language Generation Workshop, pages 33–37, Santiago de Compostela, Spain, September 4, 2017. © 2017 Association for Computational Linguistics

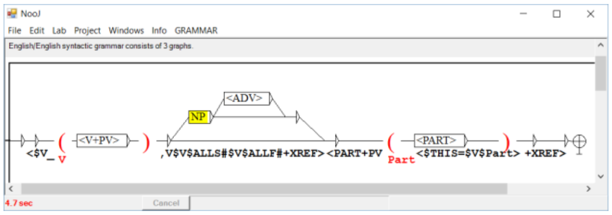
the prince ***in*** *when he arrived?* (PV)), and (3) genuine ambiguity only resolvable from context (e.g., *Her neighbor was* ***looking over*** *the broken fence*, which can mean either “looking above the fence” (preposition) or “examining the fence” (PV)). On the bright side, though, NooJ can correctly identify many discontinuous PV, such as the following:

(1) I **folded** all my bills **up** uniformly (*Great Expectations*)

(2) he had that club-hammer there ... to **knock** some one’s brains **out** with (*Moby Dick*)

(3) a program that has effectively **brought** our crime rates **down** (*Larry King Live*).

NooJ requires both a grammar and a dictionary that work in tandem to annotate PV in large corpora. Figure 1 represents an example of NooJ’s PV Grammar.



English/English syntactic grammar consists of 3 graphs.

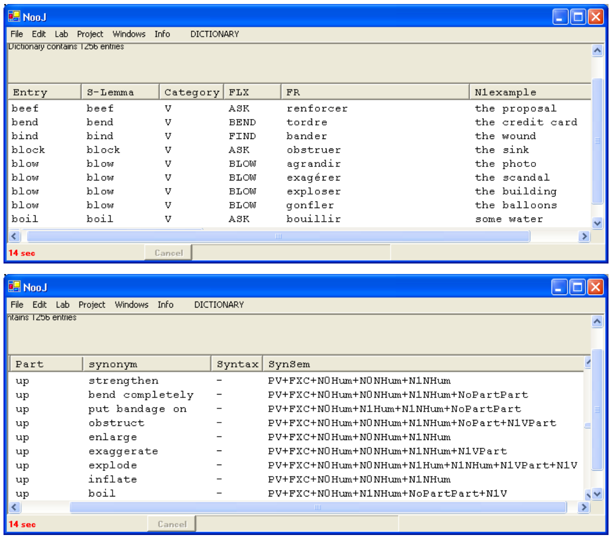
Figure 1: NooJ PV Grammar

The dictionary is based on previous work using Maurice Gross’ (1994, 1996) Lexicon-Grammar approach. Lexicon-Grammar limits abstract notions in syntax and accentuates the reproducibility of linguistic data in the form of exhaustive syntactic tables, which are manually constructed and contain both lexical and syntactic information, as can be seen in the sample Table 1. From these Lexicon-Grammar tables of PV, we created a NooJ PV dictionary that contains more than 1,200 entries, which when used in tandem with the PV grammar, could automatically annotate PV in large corpora. Figure 2 is a sample of this dictionary, which mirrors much of the information contained within the Lexicon-Grammar entry seen in Table 1.

Although early experiments identified much noise, three disambiguation grammars, adverbial and adjectival expression filters, and idiom dictionaries were added to remove false PV without creating silence. This has made for a fairly intricate way to accurately annotate PV in large corpora. Machonis (2016) explains how NooJ can successfully remove many inaccurate Text Annotation Structures (TAS).

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
+	+	beef	up	the proposal	-	+	-	-	-	strengthen
+	+	bend	up	the credit card	-	+	+	-	-	bend completely
+	-	bind	up	the wound	+	+	+	-	-	bandage
+	+	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

Table 1: Sample from PV Lexicon-Grammar


NooJ
File Edit Lab Project Windows Info DICTIONARY
14 sec
Cancel
NooJ
File Edit Lab Project Windows Info DICTIONARY
14 sec
Cancel

Dictionary contains 1256 entries

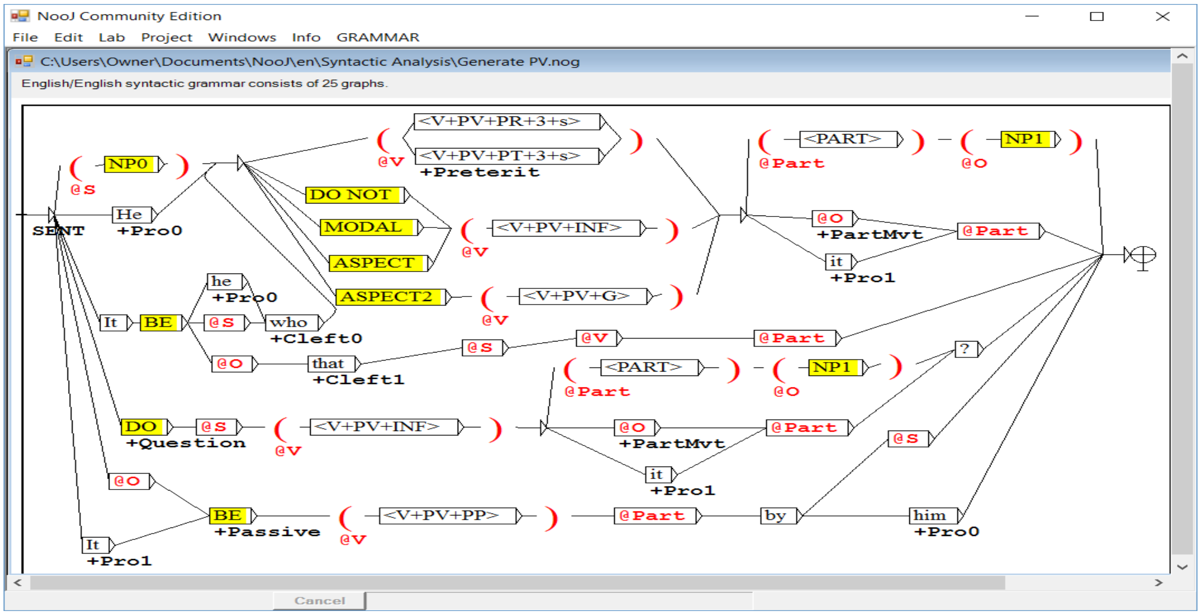
Entry	S-Lemma	Category	FLX	FR	N1example
beef	beef	V	ASK	renforcer	the proposal
bend	bend	V	BEND	tordre	the credit card
bind	bind	V	FIND	bander	the wound
block	block	V	ASK	obstruer	the sink
blow	blow	V	BLOW	agrandir	the photo
blow	blow	V	BLOW	exagérer	the scandal
blow	blow	V	BLOW	exploser	the building
blow	blow	V	BLOW	gonfler	the balloons
boil	boil	V	ASK	bouillir	some water

ntains 1256 entries

Part	synonym	Syntax	SynSem
up	strengthen	-	PV+FXC+N0Hum+N0NHum+N1NHum
up	bend completely	-	PV+FXC+N0Hum+N0NHum+N1NHum+NoPartPart
up	put bandage on	-	PV+FXC+N0Hum+N1Hum+N1NHum+NoPartPart
up	obstruct	-	PV+FXC+N0Hum+N0NHum+N1NHum+NoPart+N1VPart
up	enlarge	-	PV+FXC+N0Hum+N0NHum+N1NHum
up	exaggerate	-	PV+FXC+N0Hum+N0NHum+N1NHum+N1VPart
up	explode	-	PV+FXC+N0Hum+N0NHum+N1Hum+N1NHum+N1VPart+N1V
up	inflate	-	PV+FXC+N0Hum+N0NHum+N1NHum
up	boil	-	PV+FXC+N0Hum+N1NHum+NoPartPart+N1V

Figure 2: NooJ PV Dictionary

Overall, our NooJ PV studies have achieved 88% accuracy, 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* (preposition) vs. ***had*** *her hat and jacket* ***on*** (PV)). However, in a more recent study on the novels of Dickens and Melville, we reduced the NooJ dictionary to include only six particles (*out*, *up*, *down*, *away*, *back*, *off*) instead of twelve, which helped us achieve 98% accuracy. Other linguists, such as Hiltunen (1994:135), also limited searches to these six typical particles representing three levels of PV frequency: high (*out*, *up*), mid (*down*, *away*), and low (*back*, *off*). However, for our generation study, we used the original PV dictionary of over 1,200 entries.



C:\Users\Owner\Documents\NooJ\en\Syntactic Analysis\Generate PV.nog

English/English syntactic grammar consists of 25 graphs.

Figure 3: NooJ PV Generate Grammar

3 NooJ's PV Generating Capabilities

As an initial experiment, we created a grammar that can generate variations of a single phrase containing a PV. We changed the local variables in the original PV parsing grammar to global variables, and then added new pathways for the various transformations, following the examples in Silberztein (2015, 2016), which showed how NooJ could generate over two million transformations or parallel sentences from simple sentences, such as *Luc aime Léa* 'Luke likes Lea' and *Joe likes Lea*.

For the moment the grammar (Figure 3) only handles singular nouns in the present and past tense, but it is capable of applying a succession of transformations – particle movement, preterit, negation, clefting (both of the subject and the object), modal insertion, aspect introduction, question formation, and passive voice – to over 1,200 PV from the electronic dictionary. The upper pathway recognizes a sentence containing a PV and creates the first variants, with past tense and particle movement. So, if we enter: *Max figures out the problem*, that sentence is recognized, but NooJ also creates, *Max figured out the problem*, *Max figures the problem out*, and *Max figured the problem out*. Also, *Max* and *the problem* could be substituted by pronouns.

The next series of transformations deal with negation, modal insertion and aspect introduction. These pathways create variations such as *Max did not figure out the problem*, *Max could figure out the problem*, *Max started to figure the problem out*, *Max finished figuring out the problem*, etc. Our modal subgraph includes nine modal verbs – *can*, *could*, *may*, *might*, *must*, *should*, *ought to*, *need to*, *have to* -- as well as negative variants and contracted forms, such as *can't*, *shouldn't*, etc. Our aspectual variants include inchoative (*begin*, *start*), durative (*continue*), and completive (*finish*, *stop*), with both negative and preterit possibilities.

In the middle of the graph, we have the clefting conduit, either *It is he who*, *It is Max who*, or *It is the problem that*, followed by the same PV. In the case of clefting of the subject, the sentence will also undergo particle movement, negation, modal insertion, and aspectual variants. Thus sentences such as the following would be created: *It was Max who didn't figure it out*, *It is Max who may not figure the problem out*, *It was Max who started to figure the problem out*, *It was Max who didn't finish figuring it out*, etc.

The final two pathways involve question formation and passive voice formation: *Does Max figure out the problem?* and *The problem is figured out by Max*. These variants are also open to negative and preterit transformations, along with pronominal forms, such as: *Didn't Max figure it out?*, *The problem was not figured out by him*, etc.

Bob figured the problem out	SENT+Preterit+PartMvt
Bob has to figure it out	SENT+ModHave+Pro1
It is Bob who could figure out the problem	SENT+Cleft0+ModCan+Preterit
Mike must clear away the area	SENT+ModMust
Mike needn't clear the area away	SENT+ModNeed+Neg+Contraction+PartMvt
Mike oughtn't clear it away	SENT+ModOught+Neg+Contraction+Pro1
Brent continued to push the deadline back	SENT+Preterit+AspDurative+PartMvt
Didn't Brent push the deadline back ?	SENT+Question+Preterit+Neg+Contraction+PartMvt
It is not he who pushed the deadline back	SENT+Neg+Pro0+Cleft0+Preterit+PartMvt
Did Blake hand the exam in ?	SENT+Question+Preterit+PartMvt
He handed it in	SENT+Pro0+Preterit+Pro1
It wasn't Blake who handed in the exam	SENT+Preterit+Neg+Contraction+Cleft0+Preterit
Phil might call off the trip	SENT+ModMay+Preterit
Phil started to call the trip off	SENT+Preterit+AspInchoative+PartMvt
It isn't Phil who called it off	SENT+Neg+Contraction+Cleft0+Preterit+Pro1
Steve started turning on the radio,	SENT+Preterit+AspInchoative
Steve began to turn the radio on,	SENT+Preterit+AspInchoative+PartMvt
It wasn't Steve who didn't turn the radio on	SENT+Preterit+Neg+Contraction+Cleft0+Preterit+Neg+Contraction+PartMvt
Max couldn't burn the building down	SENT+ModCan+Preterit+Neg+Contraction+PartMvt
Max stops burning down the building	SENT+AspCompletive
The building wasn't burnt down by Max	SENT+Passive+Preterit+Neg+Contraction
Didn't Devon shred the document up ?	SENT+Question+Preterit+Neg+Contraction+PartMvt
Did Devon shred it up ?	SENT+Question+Preterit+Pro1
It isn't Devon who couldn't shred it up	SENT+Neg+Contraction+Cleft0+ModCan+Preterit+Neg+Contraction+Pro1

Table 2: Sample Sentences from NooJ PV Generate Grammar with Transformations Noted

All in all, for every sentence containing a PV, this NooJ grammar will create 2,694 entries, and sometimes more in the case of certain irregular verbs, such as *burn*, which has two past tenses, *burned* and *burnt*. This grammar can also generate sentences based on all the other PV in the NooJ PV dictionary such as, *clear away the area*, *push back the deadline, hand in the exam*, *call off the trip*, *turn on the radio*, *burn down the building*, *shred up the document*, etc. as can be seen in Table 2. If we apply the PV Generate Grammar to all of the 1,200 verbs listed in the NooJ PV dictionary, NooJ would have generated over three million different sentences.

4 Conclusion

Not only does this research shed light on a major NLG problem, i.e., generating sentences containing discontinuous multiword expressions, but it seems to approach solving the original Chomskian challenge of generating “all and only” the sentences of a language. Some of these sentences might sound more natural than others, and some will need a specific context to appear likely, however, all of the sentences generated are grammatical. Speakers may choose certain forms over others during the course of a conversation, in what might be called discourse management. Nevertheless, NooJ does allow the user to specify which transformations are to be applied and thus limit the number of sentences generated to a specific context. As can be seen in this preliminary test, NooJ is a very powerful tool for linguistics, as well as Natural Language Generation.

References

Maurice Gross. 1994. Constructing Lexicon-Grammars. *Computational Approaches to the Lexicon* (eds. Atkins and Zampolli), 213-263. Oxford University Press, Oxford, UK

Maurice Gross. 1996. Lexicon Grammar. *Concise Encyclopedia of Syntactic Theories* (eds. K. Brown and J. Miller), 244-258. Elsevier, New York.

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

Lien Huynh Hodapp. 2010. *Eliminating phrasal verbs in academic writing*. Minnesota State University Memorial Library University Archives, Mankato, Minnesota.

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. Automatic Processing of Various Levels of Linguistic Phenomena: Selected Papers from the NooJ 2011 International Conference* (eds K. Vučković, B. Bekavac, and M. Silberztein), 152-165. Cambridge Scholars Publishing, Newcastle upon Tyne, UK.

Peter A. Machonis. 2016. Phrasal Verb Disambiguating Grammars: Cutting Out Noise Automatically. *Automatic Processing of Natural-Language Electronic Texts with NooJ. Communications in Computer and Information Science book series (CCIS, volume 667),* (eds L. Barone, M. Monteleone and M. Silberztein). 169-181. Springer International Publishing AG, Cham, Switzerland.

Max Silberztein. 2015. *La formalisation des langues: l'approache de NooJ*. ISTE Editions, London,UK.

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