Look this *up* and *Try* it *out*: An Original Approach to Parsing Phrasal Verbs

Peter A. Machonis
Florida International University, Miami, Florida

Abstract

In Natural Language Processing, there has been much recent attention paid to phrasal verbs. This paper represents an attempt at parsing English phrasal verbs with and without insertion in large corpora. We used previously constructed lexicon-grammar tables of transitive and neutral English phrasal verbs, along with NooJ, a linguistic development environment that parses texts using large-scale dictionaries and local grammars. This preliminary analysis confirms that NooJ is a very useful tool not only for parsing discontinuous phrasal verbs, but also for correcting and enlarging lexicon-grammar databases. Although the overall accuracy was only 84%, the concordance generated from our most exhaustive table (the particle *up*) achieved 98% accuracy. Accuracy, as well as recall, will be enhanced in the future by adding intransitive phrasal verbs and phrasal prepositional verbs to the dictionary, as well as identifying idiomatic adverbs and other frozen expressions.

Keywords : Phrasal verbs, Lexicon-grammar, NooJ, Natural Language Processing

1. Introduction

In Natural Language Processing, there has been much recent attention paid to phrasal verbs. Linguists debate where and how to store these expressions in the lexicon, how to electronically identify discontinuous strings in a corpus, and how to avoid mistaking prepositions for particles in parsing corpora (Dehé 2002, Jackendoff, 2002, Sag et al. 2002). English transitive phrasal verbs for the most part exhibit a continuous and discontinuous form:

We **showed up** the gentleman ⇔ We **showed** the gentleman **up**

Furthermore, in the case of a pronominal direct object, only the discontinuous form is grammatical:

We **showed** him **up** ⇔ *We **showed up** him

This paper represents an attempt to parse English phrasal verbs with and without insertion in large corpora. We used previously constructed lexicon-grammar tables of English phrasal verbs, along with NooJ, a linguistic development environment that can apply a set of hundreds of lexicon-grammar entries to a large text (several million words) in real time. NooJ is freeware that can be downloaded from <http://www.nooj4nlp.net/> It requires a system that has a .NET framework. To use all of NooJ's functionalities, Windows 2000, Windows XP, or Windows Vista is also required. A 200-page manual (Silberztein 2002) is also available from the same web site.

2. NooJ: an overview

NooJ's approach to language is bottom-up. Starting with orthography, morphology and the lexicon, the basics of language are first formalized. Linguists can then use these basic levels of formalization to analyze higher and higher levels in syntax and semantics. NooJ parses texts using large-scale dictionaries and grammars, both morphological and syntactic in cascade, starting from the lowest level of formalization (character level) and moving up to structural and transformational syntax.

However, since it is not always possible to disambiguate words before performing a full syntactic and semantic analysis of a text, instead of using a tagger that obligatorily produces a certain percentage of tagging mistakes, NooJ uses a Text Annotation Structure (TAS) that holds all unsolved ambiguities (Silberztein 2007). NooJ's syntactic parser can process partially or fully ambiguous TAS. NooJ's syntactic engine uses a number of computational devices, such as finite-state transducers (to recognize multiword units) and Recursive Transition Networks (to identify frozen expressions). According to Silberztein (2002:10), NooJ's significant new feature is that it does not need specific grammars (e.g., INTEX's "meta-grammars") to parse lexicon-grammar tables; rather the grammars that are used to process frozen expressions are identical to the ones that can be used to process free sentences.

As with INTEX, NooJ can be used for simple queries, such as the following:

<V> up

which generates a concordance of all instances of verbs followed by *up* in a large corpus. In this first attempt to use NooJ for locating phrasal verbs, we examined the Henry James novel *The Portrait of a Lady*, which is included in the NooJ download. The morphological component of the grammar is fairly complete, recognizing all verb forms and tenses, and only occasionally incorrectly identifying nouns as verbs (e.g., *eyes, foot, hands, people*). These nouns, however, are interpreted as ambiguous in the NooJ generated TAS, which shows both the potential verb analysis, along with the noun reading.

Users also have the possibility of constructing their own dictionaries. We used previously constructed lexicon-grammar tables (Machonis, in press) that include an exhaustive list of transitive and neutral English phrasal verbs followed by the particle *up* (over 700 occurrences) and a rather substantial list of transitive and neutral phrasal verbs followed by other particles, such as *away*, *back*, *down*, *in*, *off*, *on*, *out*, *over*, etc. (over 500 occurrences). These tables were compiled using *The American Heritage Dictionary*, *The Longman Phrasal Verbs Dictionary*, and *The Cambridge International Dictionary of Phrasal Verbs*, along with Fraser 1976 and Spears 1996.

With the help of Max Silberztein, these tables were transferred into a NooJ phrasal verb dictionary, which was then used to identify phrasal verbs (V+PV) with simple queries such as:

<V+PV> up

<V+PV> <PART>

The first query extracts just phrasal verbs with the particle *up* while the second recognizes all phrasal verbs in the corpus.

NooJ also lets users insert pronouns (PRO) and other word forms (WF) between the phrasal verb and the particle:

<V+PV> <PRO> up

<V+PV> <WF>* up

These queries allow NooJ to extract discontinuous strings in the corpus. The first can find strings such as *burn them up, cheer me up,* etc., while the second locates strings such as *show the gentlemen up* and *sum people up*.

Upon closer examination of the NooJ-generated concordances from simple queries, however, we notice that phrasal prepositional verbs (e.g., ***look up to*** "admire", ***make up for*** "compensate") and many non-phrasal verbs, such as verb plus preposition (e.g., ***move up*** *the staircase* where you cannot say ****move*** *the staircase* ***up***) are also identified. Again, the TAS structure will indicate that *up* can not only be interpreted as a particle, but also as a preposition, verb, noun or adjective. In simple queries, NooJ furthermore recognizes some instances that have nothing to do with phrasal verbs (e.g., ***had in*** *fact occurred*, *she rested her wider* ***eyes on*** *him*), since these simple NooJ queries do not ensure that the particle is indeed associated with the verb (e.g., the expressions ***have*** (*along* + *on* + *out*) and ***eye*** *up* are in the NooJ dictionary of phrasal verbs).

3. Using a NooJ local grammar with a lexicon-grammar based dictionary

Another feature of NooJ is the possibility of constructing a local grammar or graph, which used along with the phrasal verb dictionary, can identify occurrences of phrasal verbs, both continuous and discontinuous. The advantage of using a local phrasal verb grammar is that NooJ has a rather new functionality that can associate the particle to the verb. That is, NooJ ensures that the particle (e.g., *up*) is indeed related to the verb (e.g., *look up*) in the dictionary, and thus much of the irrelevant data in the simple query concordances can be avoided. In NooJ, the grammar and the dictionary work in tandem, identifying in a text all potential instances of verb-particle combinations that are listed in the dictionary (or lexicon-grammar tables).

For the moment, our grammar is rather simple (see Figure 1), which includes an embedded NP structure that is also fairly simple. Furthermore, <P> or punctuation, is used as a delimiter after the particle following an inserted NP in order to avoid much of the noise generated when a potential particle is really a preposition.

4. Results with particle *up*

We first worked on *up*, since our lexicon-grammar tables were the most complete for this particle, which is also the most commonly occurring particle in English. Furthermore it is very frequent as a particle, rather than as a preposition, at least in *The Portrait of a Lady*, and thus shows promising results. Using the query:

<V> up

NooJ generated a concordance of 256 cases, of which only three were prepositions (e.g. *move up the staircase*), four were frozen adverbs (e.g., *pass up and down*), and three were misidentified (e.g., *foot up, people up*). It also identified many purely intransitive phrasal verbs, such as *come up, go up, glance up, grow up, jump up, loom up,* etc. which are not in the phrasal verb dictionary that is composed only of transitive and neutral phrasal verbs.

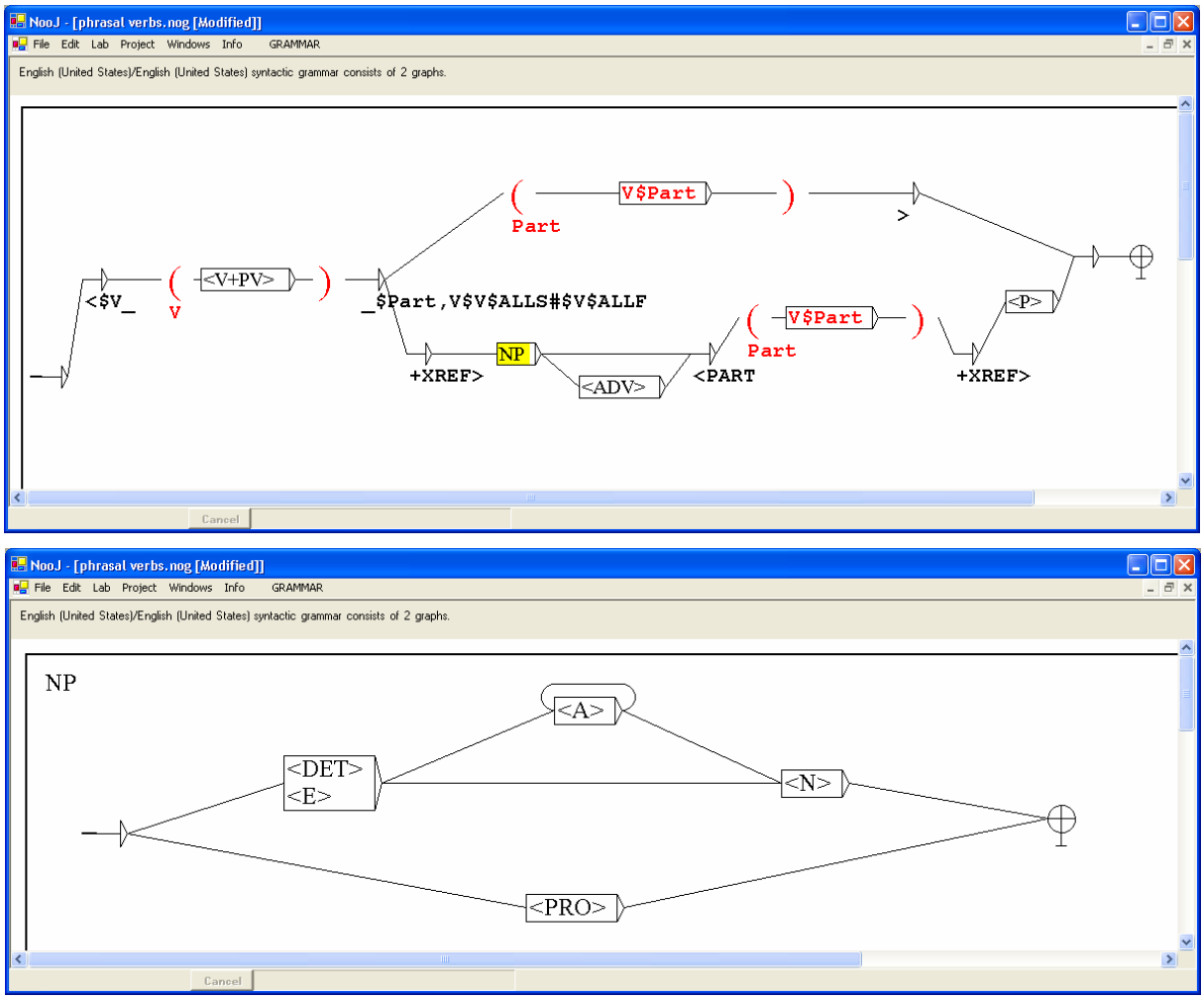

Figure 1. Phrasal verb grammar in NooJ

Neutral verbs, or verbs exhibiting neutrality (Boons, Guillet, Leclère 1976) or the causative alternation (Levin 1993), have both a transitive and intransitive linked use, such as:

I **dried up** my tears ⇔ My tears **dried up**

Whereas these potentially neutral intransitives are listed in the phrasal verb dictionary, many purely intransitive ones are not.

The query:

<V+PV> up

identified 222 occurrences, of which one was a preposition and two contained the frozen adverb *up and down*. No misidentifications occurred; however many purely intransitive phrasal verbs were missing. A comparison of these two concordances also showed some oversights in our lexicon grammar tables (e.g., *have one foot up, sit NP up, send NP up*). Thus in addition to parsing, NooJ can actually help refine dictionaries.

Next, by using the local phrasal verb grammar, along with the phrasal verb dictionary, we identified 249 cases of which six were false positives, but which included 32 discontinuous cases, all correctly identified as phrasal verbs. A sample of the discontinuous cases identified by NooJ is given in Figure 2.

hold us together or	keep	us up; marrying foreigners, forming artificial tastes, play
ll be interested in	keeping	them up." Isabel's eyes expanded as she gazed at this luri
as I say, they had	kept	it up." She was silent a little. "Why then did she want hi
Rome you'll always	look	us up. Mrs. Osmond will talk to you about the English exped
her daughter shall	make	it up." Isabel started at the words "her daughter," which
and perhaps you'll	pick	us up. I've great confidence in you; there are ever so many
r wishes. "Shall I	show	the gentleman up, ma'am?" he asked with a slightly encourag
that you're always	summing	people up." "You don't of necessity lose by that." "It's
don't say that she	sums	it all up, that would be too much to ask of her. But she su
. Yes; she likes to	take	people up. She has been very kind to me; but," she added wi
"You meant she has	taken	me up. Yes; she likes to take people up. She has been very
es as easily as she	took	them up; she worked and talked at the same time, and appear

Figure 2. Discontinuous phrasal verbs parsed from corpus (particle = up*)*

An analysis of the entire particle *up* concordance also revealed a rather thorny problem in parsing phrasal verbs, i.e., how to handle particles followed by another preposition. In most cases (26 occurrences) the phrasal verb is simply followed by a prepositional phrase. (e.g., *brought up / in Paris; get up / on the assumption; look up / with a quick glance*). In three cases, however, the verb plus the following particle and preposition must be viewed as one entity in what is known as a phrasal prepositional verb, in reality an idiomatic expression:

one **looks up to** one's brother	"admire"
that **made up for** everything	"compensate"
he's simply **wrapped up in** her	"completely absorbed in"

Phrasal prepositional verbs such as these will also have to be added to the lexicon-grammar and compiled into a dictionary in order to achieve an accurate parsing of English.

On the other hand, some ambiguity seems inevitable. For example, in our NooJ generated concordance of *The Portrait of a Lady*, the expression *look up to* appears as a phrasal prepositional verb as above, but also as a correctly parsed phrasal verb followed by an infinitive (*look up / to hear*) as in the following example:

she **looked up** to hear Lord Warburton announcing that he should like to marry

5. *Out* and other particles

Like the particle *up*, *out* seems to occur most frequently as a particle: out of 107 entries in the *out* concordance, only four cases were not particles: the only prepositions parsed by NooJ were examples of *out of* and the two incorrect occurrences in the concordance were *have found out* and *spend the evening out*, since the expressions *have something out* "have removed" and *even something out* "make equal" are part of the phrasal verb dictionary. A small sample of the discontinuous cases involving *out* generated by NooJ is given in Figure 3.

erests; but I can't	make	them out." Ralph leaned back in his chair with folded arms
e; I forgot you had	pointed	that out. Of course," Madame Merle added, "you've had infin
dull people always	put	him out; but a quick and cultivated girl like Isabel would
houldn't attempt to	reason	it out--you never know where it may lead you. There are som
situation. She had	reasoned	the matter well out, making it perfectly clear that she bro
t I shall certainly	take	it out. Ralph, of course, has Gardencourt; but I'm not sure
pparently, had been	thinking	the thing out, but had arrived at a different conclusion fr
l, you seem to have	thought	it out," said Mr. Touchett. "But I don't see why you appeal
As I tell you, I've	turned	him out." "Yes; but a lover outside's always a lover. He's
very opportunity of	turning	him inside out." "Well, he may do for one letter, but what

Figure 3. Discontinuous phrasal verbs parsed from corpus (particle = out*)*

Although the data in our lexicon-grammar tables for other particles was less comprehensive, we nevertheless transferred them into the phrasal verb dictionary and applied the phrasal verb grammar to see if NooJ could correctly identify other English particles such as *about, along, away, back, down, in, off, on, over* and *through* in our corpus. The results of this preliminary parsing are given in Figure 4, along with the results of our parsing of *up* and *out*.

While some of the particles presented minor problems, one major disappointment concerned the particle *in*. Out of 84 cases, only 18 were correctly identified as particles, which did include five discontinuous cases (i.e., *you keep bringing her in, he drank it all in, and locked herself in, while taking it in, Osmond however took him in*). The rest were clearly prepositions, of which 22 were prepositions preceded by nouns incorrectly associated with phrasal verbs, such as in the following example: *a sudden break in her voice*. This happens since the lexicon-grammar contains the following phrasal verbs: *break something in, buckle someone in, clock someone in, hand something in, time someone in*.

Although less frequent in our corpus than the particle *in*, the preposition *on* also was mistakenly parsed as a particle in 50% of the cases. In many of the cases, however, if frozen expressions were first identified, this problem could be largely eliminated. In fact half of the prepositional usages of *on* involved idioms: *on DET occasion, have on POSS-0 mind, on the whole*. Likewise, factors such as pronouns appearing after the particle could also eliminate mistakenly identified particles such as: *on him, through us*.

On the other hand, just about all examples of *about, away, back, down, off*, and *over* were correctly parsed. The only exceptions were: *run down to Spain* and *keeps coming back*. This last example confirms the importance of identifying all intransitive uses of phrasal verbs (e.g., *to come back*) before parsing transitive ones (e.g., *to keep (someone + something) back*).

6. Conclusions

Although a first attempt at parsing phrasal verbs with NooJ, this original linguistic development environment seems extremely promising – not only for parsing phrasal verbs, but also for correcting and enlarging lexicon-grammar databases. In addition to showing the importance of adding intransitive phrasal verbs and phrasal prepositional verbs to the

dictionary, as well as identifying idiomatic adverbs and other frozen expressions first, this preliminary analysis confirms that listing all phrasal verbs in a lexicon-grammar database can help eliminate confusing prepositional usages with true particles.

Particle	# of examples	Correct continuous examples	Correct discontinuous examples	Prepositions	Misidentifications	Percentage of incorrect	Percentage of correctly identified phrasal verbs
about	3	3				0.00%	100.00%
along	1			1		100.00%	0.00%
away	14	10	4			0.00%	100.00%
back	20	16	3		1	5.00%	95.00%
down	30	25	4	1		3.33%	96.67%
in	84	12	6	27	39	78.57%	21.43%
off	35	24	11			0.00%	100.00%
on	20	10		10		50.00%	50.00%
out	107	84	19	2	2	3.74%	96.26%
over	15	9	4	2		13.33%	86.67%
through	5	1	1	3		60.00%	40.00%
up	249	211	32	3	3	2.41%	97.59%
TOTAL	583	405	84	49	45	16.12%	83.88%

Figure 4. Percentage of correctly identified phrasal verbs

In addition to improving accuracy, we are also working on expanding coverage (recall) of all phrasal verbs in a given text. By eliminating the punctuation node of the graph and in expanding the NP node, NooJ did locate 60 more valid discontinuous cases of verb particle combinations in *The Portrait of a Lady*. However, the overall accuracy was much lower: only 68% accuracy as opposed to 84% accuracy with the simple NP node and punctuation node. As further improvements are achieved with NooJ, for example building a syntactic parser to recognize sentences, recall and accuracy should both improve. Likewise, expanding the coverage of our phrasal verb dictionary should help.

Although many linguists shun the idea of listing every verb-particle combination in a dictionary, claiming the value of generalization and avoidance of redundancy, not all verbs can combine with particles, even when associated with a productive semantic class. Villavicencio (2005) in investigating certain of Levin's (1993) classes of verbs in random combination with six particles has shown that some of Levin's classes – such as "Clear" verbs and the "Bring and Take" class of verbs – show a high degree of productivity, while others do not. Although Villavicencio (2005:430) argues that this information could be stored as lexical redundancy rules for the more productive classes, and listed in the lexicon for the non-

productive classes, examining more closely the most productive classes might be yet another way of expanding our phrasal verb dictionary.

Nevertheless, certain problems with ambiguity will remain. As we have seen in concordances generated by NooJ, expressions such as *look over* and *move along* are extremely difficult to parse correctly, i.e., to distinguish particle from prepositional usages:

The doctor looked (her shoulder over + over her shoulder) "examine"
The neighbor looked (*her shoulder over + over her shoulder) "look" plus preposition

They moved (the children along + along the children) "ask to leave"
They moved (*the platform along + along the platform) "move" plus preposition

Expanding selectional restrictions of complements in the lexicon-grammar will clearly help solve the issue of ambiguity, but a full syntactic and semantic analysis of a text might not be enough. Some type of discourse analysis at the pragmatic level is the only way to fully distinguish these complex ambiguities.

References

American Heritage Dictionary of the English Language, (2000), Houghton Mifflin, Boston.

BOONS J-P., GUILLET A., & LECLÈRE C. (1976), *La structure des phrases simples en français, Constructions intransitives*, Droz, Geneva.

Cambridge International Dictionary of Phrasal Verbs, (1997), Cambridge University Press, Cambridge. DEHÉ N. (2002), *Particle verbs in English, Syntax, information structure and intonation,* John Benjamins, Amsterdam/Philadelphia.

FRASER B. (1976), *The verb-particle combination in English*, Academic Press, New York.

JACKENDOFF R. (2002), "English particle constructions, the lexicon, and the autonomy of syntax", in DEHÉ N., JACKENDOFF R., MCINTYRE A. & URBAN S. (eds.), *Verb-particle explorations*, Mouton de Gruyter, New York: 67-94.

LEVIN B. (1993), *English verb classes and alternations: a preliminary investigation*, University of Chicago Press, Chicago.

Longman Phrasal Verbs Dictionary, (2000), Pearson Education Limited, Essex.

MACHONIS P. (in press), "Disambiguating Phrasal Verbs", in *Lingvisticae Investigationes*.

SAG I., BALDWIN T., BOND F., COPESTAKE A., & FLICKINGER D. (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, Mexico City: 1-15.

SILBERZTEIN M. (2002), *NooJ Manual,* <http://www.nooj4nlp.net/>

SILBERZTEIN M. (2007), "An Alternative to Tagging", in *Proceedings of NLDB 2007, Lecture Notes in Computer Science*, Springer Verlag, Berlin: 1-11.

SPEARS R. (1996), *Basic Phrasal Verbs*, NTC Publishing Group, Lincolnwood.

VILLAVICENCIO A. (2005), "The availability of verb-particle constructions in lexical resources: How much is enough?", in *Computer Speech and Language* 19: 415-432.