

A Construction Grammar Approach to Prepositional Phrase Attachment: Semantic Feature Analysis of V NP1 *into* NP2 Construction*

Liyin Chen^a, Siaw-Fong Chung^a, and Chao-Lin Liu^b

^aDepartment of English, National Chengchi University,
NO. 64, Sec 2, ZhiNan Rd., Wenshan District, Taipei City 11605, Taiwan
{98551505, sfchung}@nccu.edu.tw

^bDepartment of Computer Science, National Chengchi University,
NO. 64, Sec 2, ZhiNan Rd., Wenshan District, Taipei City 11605, Taiwan
chaolin@nccu.edu.tw

Abstract. This paper provides a construction grammar perspective to identifying the ambiguity of prepositional phrase (PP) attachments (i.e., whether a PP is attached to the closest VP or NP1). Despite the wide discussion of these two structures (VP-attached and NP1-attached), we raise the possibility for a third parsing structure (about 11.3% from all 1845 instances analyzed), a co-attachment to both verb and noun. A co-attachment structure denotes the lack of [movement] feature in both the verb and noun surrounding a PP. This proposal is arrived when we annotate the semantic feature [-movement] to both VP and NP1, respectively, in a caused-motion construction of V NP1 *into* NP2 (e.g., *vote an individual into the presidency; shamed us into pity; define ourselves into a box*).

Keywords: construction grammar, prepositional phrase attachment, caused-motion construction.

1 Introduction

The preposition *into* describes the path of motion event which typical involves an object, or figure, moves along the path to enter a reference object, or ground (Talmy, 2000). An example of motion event is the caused-motion construction involving a verb (V) and two noun phrases (NP1 and NP2) as a direct and an indirect object, respectively. Sentence (1), extracted from the Penn Treebank Wall Street Journal (WSJ) Corpus¹ (Charniak, et al., 2000), illustrates such a V NP1 *into* NP2 construction (shown in bold with lexical categories glossed underneath). The basic semantics of the construction involves a motion event that requires the direct object (NP1) to be moved and directed to the confinement of indirect object (NP2). In this case, an unspecified number of airplanes undergo movement towards a deictic space.

- (1) **To shove even more airplanes into this space** is asking for trouble,
 V NP1 Prep NP2
experts say. (WSJ-V1141)

However, this type of prepositional phrases poses an ambiguity problem in parsing. Sentence (1) serves as an example for one means of parsing in which the preposition closely associates with the verb but not NP1. The second possibility of parsing is where the PP is required to be interpreted with NP1, as illustrated in bold in sentence (2).

* The authors would like to thank the anonymous reviewers for their comments on this work. This study is supported in part by National Science Council Research Grants (NSC 99-2410-H-004-206-, NSC 100-2628-H-004-137-, and NSC-100-2221-E-004-014).

¹ All the examples discussed in this paper are from the WSJ corpus unless specified.

- (2) *And he soon became aware that the government was able
to show a flow of millions of dollars in illicit funds into his account.*
 V NP1 Prep NP2

Here the head of NP1 (*a flow*) is to be interpreted along with *into* and NP2 (*his account*), rather than with the preceding verb (*to show*). Computational linguists have found these two structures causing parsing problems in natural language processing (NLP) and referred to this problem of determining the site of PP to be attached as the PP attachment problem (e.g., Hindle and Rooth, 1993; Volk, 2006). As illustrated in (1) and (2), this problem is conventionally formalized as a binary choice (Merlo and Ferrer, 2005), either verb-attached for (1) or noun-attached for (2). In the minimalist syntax, ternary structures like (3a) are to be transformed by deriving an explicit causative construction (3b) (Radford, 2004). The operation involves raising of the verb *roll* to join the causative verb *made* to adhere to a binary operation.

- (3a) *He rolled the ball down the hill.*
 V NP1 Prep NP2
- (3b) *He made + roll the ball (roll) down the hill.*
 V-causative + V NP1 trace Prep NP2
 (Radford, 2004, p. 337, with gloss added)

Although the plausibility of equating the two constructions has long been questioned (e.g., Fodor, 1970), the causative structure (3b) cannot provide a direct solution to the PP attachment problem for (3a). Moreover, the binary solution to the problem has been challenged by computational linguists. For example, Merlo and Ferrer (2005) contend that such a dichotomous treatment may be a simplification. They propose to take into account of the nature of the attachment by distinguishing PP arguments from PP adjuncts. Sentence (4) is an example of two verb-attached PPs that maintain different relationships with the verb shown in the gloss.

- (4) *Put the block on the table in the morning.*
 V NP1 PP argument PP adjunct
 (Merlo and Ferrer, 2005, p. 342, with gloss added)

Since PP arguments carry the core message and PP adjuncts provide additional information to the core meaning, their distinction further refines NLP tasks. Although studies like Merlo and Ferrer (2005) provide novel approaches to tackle the PP attachment problem, the notion of binary sites for PP attachment has not been scrutinized. The presupposition of binary attachment sites, however, may result in a forced selection from one of the two choices and may overlook other possibilities for correct parsing. Consider the construction in bold in sentence (5) for determining the PP attachment site.

- (5) *Frank sneezed the tissue off the table*
 V NP1 Prep NP2
 (Goldberg, 1995, p. 152, with gloss added)

According to our first choice, verb-attached parsing, the verb *sneezed* is to be analyzed with NP1 *the tissue*. The grouping is semantically invalid since the verb is normally intransitive without a direct object. Yet, it is not any less awkward as the noun-attached option is considered (*the tissue off the table*). In Goldberg's (1995) seminal work on construction grammar, she discusses the basic semantics of caused-motion construction or that "the causer argument directly causes the theme argument to move along a path designated by the directional phrase: that is, 'X CAUSES Y to MOVE Z'" (p. 152). In brief, the caused-motion construction includes a directional phrase like *into* PP and entails a movement feature. However, (5) illustrates an atypical example of caused-motion construction where the construction fails to be interpreted through its components or what the PP attachment problem is based on. According to Goldberg, the semantic meaning of (5) can only be derived by taking into account of the entire construction. In other words, to address the PP attachment issue in sentences like (5), we need to

take into account of a third possible structure in addition to a binary choice from verb- or noun-attachment.²

In this study, we take the construction grammar approach to reformatize the PP attachment problem. In addition to the conventional binary approach to determining the PP attachment sites, we suggest a third possible structure where the PP co-attaches to both verb and noun based on the construction grammar framework. We also develop a semantic analysis of the feature movement (denoted as [+movement] or [-movement]) for the verb and direct object in the V NP1 *into* NP2 construction to determine the PP attachment site. Our proposal examines the WSJ corpus data by means of manual annotation.

2 A Semantic Feature Classification

From our preliminary observations of the WSJ data, we noticed that the notion of movement is closely correlated to the verb and first noun in the V NP1 *into* NP2 construction, and the observation can be confirmed by the semantic meaning of the preposition *into*. According to Tyler and Evans (2003), the ‘proto-scene’, or illustration of the primary sense in spatial configuration, of *into* involves NP1 (figure) being outside NP2 (ground) and then entering the boundary of NP2. The movement feature is a distinctive feature for *into* to be distinguished from *in* as the latter expresses a locative sense. As the notion of path is encoded by the preposition, the parsing structure may reflect our proclivity to associate the notion of path with that of movement. We therefore hypothesize that the feature [\pm movement] can be used to determine the PP attachment site as summarized in Table 1.

Table 1: Hypotheses for determining PP-attachment sites from [\pm movement] specification at V and NP1.

[\pm movement]		Hypotheses
V	NP1	
+	-	Expected specification for verb-attached PP structure.
-	+	Expected specification for noun-attached PP structure.
-	-	Expected specification for co-attachment structure.
+	+	Undetermined.

Table 1 demonstrates the four possible outcomes after specifying [\pm movement] at V and NP1, each elaborated below. Constructions (in bold) in (6) and (7) exemplify verb and noun attachment, respectively, and the feature is specified in plus or minus sign for each lexical item in gloss.

(6) *They threatened **to crash the jet into Kuwait's royal palace.*** (WSJ-V924)
 V [+] NP [-] PP

(7) *He **describes his launch into American society** In 1962,...* (WSJ-V245)
 V [-] NP [+] PP

It is not difficult to see that the concept of path encoded in the PPs is more semantically coherent when interpreted with [+movement] sites such as the verb *crash* in (6) and the noun *his launch* in (7). In contrast, when the potential attachment site is [-movement] like the NP *the jet* in (6) or the verb *describe* in (7), there is no strong semantic association with the PP. However, the binary distinction of verb and noun attachment sites is not sufficient to capture the parsing of

² Such treatment appears to be in common with approaches that accommodate ternary branching such as Head-Driven Phrase Structure Grammar or HPSG (Pollard and Sag, 1994), or proponents like Jackendoff (2002) and Carrier and Randall (1992). In contrast to the other approaches, this study stresses the entirety of construction in which causative meaning of non-causative verbs can be derived as in the case of *sneezed* in (5). In addition, much earlier work concentrated on the analysis of ternary branching resultative constructions (e.g., Carrier & Randall, 1992; Wechsler & Noh, 2001). To the best of our knowledge, the results have not been applied to tackling the PP attachment problem.

some instances like (8). Sentence (8) exemplifies an unconventional caused-motion construction like those described by Goldberg (1995).

- (8) *Watching the weekend spectacle, we return to the thought that the next U.S. President and the rest of the West's leaders should keep in mind that their primary responsibility is not to baptize the Soviet Union into Western civilization.* (WSJ-V30)
- V [-] NP [-] PP

The verb *baptize* is commonly used in passive voice and is an unaccusative verb with its object as the argument. Yet, in sentence (8) we found a transitive use followed by a direction prepositional phrase. A semantic feature of movement can only be construed when the entire construction V NP1 *into* NP2 is processed simultaneously. This is the basis of our contention that in addition to previous two sites to attach PP, a co-attachment structure is required.

The last feature category [+ , +] is the most complicated one among the four. Consider sentences (9) and (10). Complication arises as the intensity of [+movement] could vary and compete between the verb and noun that would result in a differential reading. For example, sentence (9) demonstrates a stronger sense of movement in the verb that results in a verb-attachment interpretation. On the contrary, sentence (10) posits a stronger feature in the noun that results in a noun-attached reading. In addition, the rate of V-NP or NP-PP occurrence may also interfere with the feature reading.

- (9) *But the mess in sales isn't the only problem that slowed revenue growth into single digits in the latest quarter.* (WSJ-V1334)
- V [+] NP [+] PP

- (10) *Moreover, the economy shows scant evidence of repeating the post-1929 slide into the Depression.* (WSJ-V1337)
- V [+] NP [+] PP

However, the determination of feature strength is subjective which requires a more objective approach such as cognitive experiments to be conclusive; therefore, we designate this category as 'undetermined'.

3 Categorization Procedures

The extracted WSJ corpus data were analyzed manually to identify the semantic category for all words in the lexical categories, V and NP1. According to Fellbaum (1990), there are 15 files or semantic domains of verbs in WordNet all of which are listed in Table 2 except for *weather*. The online WordNet Search 3.1 was used to identify the category of each verb by selecting "show all" in Display Options for each search word (Figure 1). The categorization fits our purpose of identifying the presence or absence of the semantic feature [\pm movement]. We based our judgment from the senses listed in the Oxford Online Dictionary (<http://oxforddictionaries.com/>) for *into*, except for the addition of two domains, *body* and *possession*, in Fallbaum. Table 2 summarizes the semantic domains that reflect the [\pm movement] feature with examples. In Figure 1, the verb *induct* is used as an example to show the display of search results. While *induct* has at least five senses, the senses fall into three semantic domains, as designated in < >, namely, *social*, *creation* and *communication*. For each instance from the WSJ corpus, the sense of the verb was identified first by one of the authors and then matched to the WordNet search results. For example, the meaning of *induct* in sentence (11) is *introduce*, thereby assigning *communication* (sense five above) for categorization.

- (11) *Typically, Mr. Grace inducts his Catholic recruits into the Knights of Malta and encourages their participation in charity.* (WSJ-V855)

However, if the WordNet result did not have a match with the verb under search, dictionaries (e.g., Merriam-Webster online dictionary and Collins English Dictionary) were consulted. In total, there were 30 verbs (1.6%) (mainly words with prefixes such as *reopen* and *unimpede*) consulted dictionaries for their categories.

Table 2: Semantic Categories of Verbs (Fallbaum, 1990).

[+movement]	Central Notions (with examples from WSJ)	[-movement]	Central Notions (with examples from WSJ)
Motion	Make a movement or contained motion e.g., <i>fall, elevate, intrude</i>	Cognition	Cognitive acts and states. e.g., <i>await, engineer, trick</i>
Contact	Troponyms of fasten, cover, cut and touch e.g., <i>chisel, nudge, pile</i>	Communication	Verbal and non-verbal e.g., <i>say, cite, protest</i>
Change	Verbs of becoming. e.g., <i>substitute, accelerate, delay, empty, restructure</i>	Consumption	Verbs of ingesting, using, exploiting, spending and sharing e.g., <i>drink, wean, tax</i>
Creation	By mental act, artistic means or from raw material e.g., <i>rebuild, cast, cause</i>	Social	Include law, politics, economy, education family and religion. e.g., <i>resign, punish, hinder</i>
Competition	Includes sports, games and warfare. e.g., <i>battle, parlay, fight</i>	Perception	Covers the five senses. e.g., <i>find, witness, see</i>
Body	Bodily care and function words. e.g., <i>breathe, sweat, hypnotize, secrete, twist</i>	Emotion	Subject or object as the experiencer. e.g., <i>scare, tease, rile</i>
Possession	Change of possession and its preceding and resultative state. e.g., <i>buy, invest, spend</i>	Stative	Verbs of being and having. e.g., <i>require, defer, involve</i>

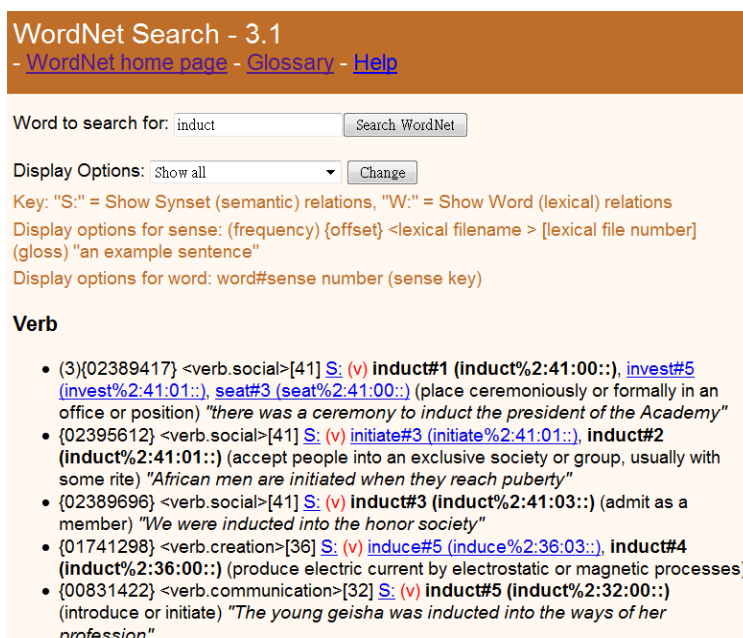


Figure 1: A screenshot of WordNet Search 3.1 search result for the verb *induct*.

The categorization of nouns in WordNet unfortunately does not focus on the notion of movement as this feature is not a primary characteristic of noun. Different criteria were used for

categorizing nouns. The nouns in our constructions were grouped manually into six major categories. The [+movement] feature is found in the following three domains: movement (e.g., verb derivations like *continuation*, *delivery* and *expansion*), route/entry (e.g., *way*, *window*) and psychological/cognitive path (e.g., *insight*, *inquiry*, *research*). Most of the [+movement] nouns were easily identified as they are closely related to their morphological counterparts in the verb category (e.g., *discharge*, *investigation*, *immigration* and *investment*). The [-movement] nouns can be found under another three domains: entity (e.g., physical objects), abstraction (e.g., temporal and spatial concepts), group/person (e.g., name entities and pronouns).

4 The Corpus

The annotated Penn Treebank WSJ corpus was used for data analysis. Two sub-corpora that contain the V NP1 *into* NP2 were extracted based on the built-in annotation: verb-attached structure ((VP NP1)(*into* NP2))³ and noun-attached structure ((VP)(NP1 *into* NP2))⁴. All the data were converted into Excel files for further processing. First, those instances without an NP (labeled as ‘None’ in WSJ) were excluded. Second, to provide a full coverage of verbs, only one instance per lexeme was selected for manual semantic analysis. For example, there are in total 29 instances of the V NP1 *into* NP2 construction for the lemma *absorb* (four for *absorb*, 22 for *absorbed*, and three for *absorbing*). Only one instance of each lexical form (*absorb*, *absorbed* and *absorbing*) was analyzed. In addition, instances which do not have a PP attached to either the verb or head noun in NP1 were manually eliminated. Sentence (12) exemplifies such a case as *into* is neither attached to V (*sympathizes*) nor NP1 (*a bit*), but rather to another NP (*Belushi’s escape*) subsumed under NP1.

(12) *If she was as simpering in life as she is on film, one sympathizes a bit with Belushi’s escape into reality-altering substances.* (WSJ-V17)

In the end, a total of 1918 types of verbs were screened, and 1845 instances were considered for further analysis along with their direct object (NP1).

5 Annotation Results

The distribution of the four types of [\pm movement] feature specified at V and NP1 is summarized in Table 3. Overall, over half of the PPs are attached to the verb [+ , -] (51.5%). Noun-attached [+ , -] and undetermined structures [+ , +] fall into close range (17.4% and 19.8%, respectively). The lowest proportion is co-attachment [- , -] (11.3%) which is expected due to its atypical semantic properties.

Table 3: Distribution of four types of feature specification.

Hypothesized attachment site	[\pm movement]		Rate (count)
	V	NP1	
Verb-attached	+	-	51.5% (950)
Noun-attached	-	+	17.4% (321)
Co-attached	-	-	11.3% (208)
Undetermined	+	+	19.8% (366)
Total			100% (1845)

Next, we compare our annotation results with that extracted directly from the Penn Treebank which was generated from the traditional binary attachment approach. Table 4 demonstrates the feature distribution of the Penn Treebank binary attachment in percentage and counts in parentheses. A match of 65.6% for verb-attached PPs and 44.8% for noun-attached PPs can be

³ An example of noun-attached structure: (VP (VBZ parses) (NP#1230 (NNS names)) (PP-CLR (IN into) (NP (DT every) (JJ conceivable) (NN interest) (NN group))))).

⁴ An example of noun-attached structure: (VP (VB plot) (NP (DT a) (JJ peaceful) (NN course)) (PP-CLR (IN into) (NP#1080 (DT the) (NN future))))).

found across both methods. Following the feature analysis, 137 instances of noun-attached PPs are found under the verb-attached category in the Penn Treebank annotation. More than half of these instances are communication and cognitive verbs (100/137). Examples include *masterminded*, *consolidated*, *picked*, *implies*, *mandate*, *requested* and *protested*. For those verb-attached PPs classified under the noun category, the main verb categories are change and motion verbs (39/76), exemplified by verbs like *fermented*, *infusing*, *integrating*, *transforming*, *reached*, *walk*, *sweeping*, *threw* and *drops*.

Table 4: Feature analysis of the Penn Treebank computer-generated binary attachment.

Hypothesized attachment site	[± movement]		Penn Treebank Binary Attachment	
	V	NP1	Verb-attached PPs	Noun-attached PPs
Verb-attached	+	-	60.9% (874)	18.5% (76)
Noun-attached	-	+	9.6% (137)	44.8% (184)
Co-attached	-	-	14.2% (204)	1.0% (4)
Undetermined	+	+	15.3% (219)	35.8% (147)
Total			100% (1434)	100% (411)

A further analysis of the semantic domains of all the co-attached PPs can be found in Table 5. The social, emotion, cognition and communication domains comprise of nearly 90% of the data.

Table 5: The distribution of semantic domains of verbs in [- -] category.

Verb domains	Communication	Social	Emotion	Cognition	Stative	Consumption	Perception	Total
Count	94 (45.2%)	48 (23.1%)	31 (14.9%)	16 (7.7%)	9 (4.3%)	6 (2.9%)	4 (1.9%)	208 (100%)

Both the social and communication domains consist of verbs that can only be interpreted with rhetorical force as shown bolded in (13a) and (13b), respectively.

(13a) ...***punish Iran and Iraq into an agreement*** on each other's production quotas... (WSJ-V42)

(13b) ...*the courts have refused to uphold contracts in which people have voluntarily **contracted themselves into peonage or slavery.*** (WSJ-V30)

Emotion verbs are also found to have their emotional sensation evoked and passed onto NP2 as shown bolded in (14). Some of them collocate with *into* (e.g., *intimidate*, *scare*, and *galvanize*). This domain is almost uniformly co-occurs with human subjects (93%).

(14a) *It happened in the 1970s when the government **panicked itself into an "energy crisis,"*** (WSJ-V202)

(14a) ...***deluded ourselves into thinking*** we were safe. (WSJ-V70)

In contrast, the result for cognition verbs is rather difficult for interpretation as the category could have covered too broad a range. Verbs like *trick*, *sorted*, *plugged*, *instilling*, *reclassify*, *parsing*, and *categorizing*, clearly denote change at the cognitive level and are supposedly assigned to the verb-attached group. In brief, the above annotation results show that the co-attached instances for the *into* PP are more commonly associated with interpersonal functions such as communication, social and emotion.

6 Conclusion

In this study, we adopt the construction grammar framework to provide a different means to reformulate the PP attachment problem. In addition to the conventional approach that makes a

binary choice between verb- and noun-attached sites, we propose a third possible parsing structure that requires a co-attachment to both the verb and noun. By exploiting the linguistic properties of caused-motion construction, we propose the use of the semantic feature [movement] to parse tree structure. The co-attached structure lacks the [movement] feature at both V and NP1 because the sense of motion resides in the construction per se rather than being imposed on the attachment sites. The results indicate that 11.3% (208 instances) of the 1835 types of verbs in V NP1 *into* NP2 constructions extracted from the WSJ corpus are co-attached. It is therefore worthy of further consideration in NLP tasks involving PP-attachment.

However, there are some limitations to the feature specification approach of this study. First, more stringent criteria for feature annotation are necessary. For example, some words in communication, cognition and social interaction domains denote rhetorical forces (e.g., *entice*, *allure*, *pressure*) and their movement feature may have been overlooked. Furthermore, refinement on the undermined category [+ , +] is necessary to provide more accurate figures to support our approach. Future work should also include analysis of the nouns in depth, and extend the results of this study to other prepositions and PPs in other constructions.

References

- Carrier, J. and J. Randall. 1992. The argument structure and syntactic structure of resultatives. *Linguistic Inquiry*, 23(2), 173–234.
- Charniak, E., D. Blaheta, N. Ge, K. Hall, J. Hale and M. Johnson. 2000. *BLLIP 1987-89 WSJ Corpus Release 1*. Linguistic Data Consortium, Philadelphia
- Fellbaum, C. 1990. English verbs as a semantic net. *International Journal of Lexicography*, 3(4), 278-301.
- Fodor, J. 1970. Three reasons for not deriving “kill” from “cause to die”. *Linguistic Inquiry*, 1(4), 429–438.
- Goldberg, A. 1995. *Constructions: A construction grammar approach to argument structure*. Chicago: The University of Chicago Press.
- Hindle, D. and M. Rooth. 1993. Structural ambiguity and lexical relations. *Computational Linguistics*, 19(1), 103-120.
- Jackendoff, R. 2002. *Foundations of language: Brain, meaning, grammar, and evolution*. Oxford: Oxford University Press.
- Litkowski, K. and O. Hargraves. 2007. Sem-Eval-2007 Task 06: Word-sense disambiguation of prepositions. *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pp. 24-29.
- Merlo, P. and E. E. Ferrer. 2006. The notion of argument in prepositional phrase attachment. *Computational Linguistics*, 32(3), 341-377.
- Radford, A. 2004. *Minimalist syntax: Exploring the structure of English*. Cambridge: Cambridge University Press.
- Pollard, C. and I. A. Sag. 1994. *Head-driven phrase structure grammar*. Chicago, IL: University of Chicago Press.
- Talmy, L. (2000). *Toward a cognitive semantics. Vol. II: Typology and process in concept structuring*. Cambridge, MA: The MIT Press.
- Tyler, A. and V. Evans. 2003. *The semantics of English prepositions*. New York: Cambridge University Press.
- Volk, M. 2006. How bad is the problem of PP-attachment? A comparison of English, German and Swedish. *Proceedings of the Third ACL-SIGSEM Workshop on Prepositions*, pp. 81-88, Trento, Italy, April 2006.
- Wechsler, S. and B. Noh. 2001. On resultative predicates and clauses: Parallels between Korean and English resultatives. *Language Sciences*, 23, 391-423.