Title	Computational Parsimony in the Case of V-V Compounds in Japanese
Author(s)	橋本　力（HASHIMOTO Chikara）
Citation	Theoretical and applied linguistics at Kobe Shoin, No.8：37-51
Issue Date	2005
Resource Type	Bulletin Paper / 紀要論文
Resource Version	
URL	
Right	
Additional Information	

Computational Parsimony in the Case of V-V Compounds in Japanese*

HASHIMOTO Chikara

Abstract

As a type of Multiword Expression (Sag et al., 2002; Baldwin & Bond, 2002), Japanese verbal compounds (V_1-V_2 compounds, hereafter) pose serious problems for Japanese Natural Language Processing (NLP) and require a sophisticated linguistic treatment. Hashimoto (2004) presented such a treatment of V_1-V_2 compounds based on the JACY framework (Siegel & Bender, 2002). However, the treatment suffered from overgeneration involving what Matsumoto (1996) calls 'V_1-V_2 with semantically deverbalized V_1', which is peculiar in that, even though it shows (partial) compositionality, its productivity is very restricted. In this paper, I propose an alternative analysis for V_1-V_2 with semantically deverbalized V_1, where all V_1-V_2s of that kind are regarded as single words to account for their restricted productivity but are given (partially) compositional semantic representations. Also, I report on an evaluation experiment that shows an advantage of the alternative treatment. Finally, I argue that grammar developers should take into account computational parsimony; we should not try too hard to generalize phenomena that we can easily enumerate exhaustively.

1. Introduction to the Problem of V_1-V_2 Compounds

Although recent Natural Language Processing (NLP) systems have relied mainly on shallow processing techniques, some NLP problems still need a deep linguistic treatment. Among such problems is the one that is brought about by **Multiword Expressions (MWEs)** (Sag et al., 2002; Baldwin & Bond, 2002). Sag et al. (2002) and Baldwin and Bond (2002) define MWEs as "idiosyncratic interpretations that cross word boundaries (or spaces)" and illustrate the problem of MWEs at length. In Hashimoto (2004), I regarded V_1-V_2 compounds in Japanese as a type of MWE. As such, they pose vexing problems for Japanese NLP.

In Japanese, which is an agglutinative language, V_1-V_2 compounds abound in both spontaneous speech and written documents, and their surface compositions are quite simple: an infinitive verb followed by another verb. However, their usages and meanings are so complex that they have been one of the central issues of Japanese linguistics (Teramura, 1969; Yamamoto, 1983; Tagashira & Hoff, 1986; Kageyama, 1993; Matsumoto, 1996; Himeno, 1999; Fukushima, 2003).

Some V_1-V_2s are productive and transparent in their meanings, while others show highly lexicalized characteristics. Below are examples of V_1-V_2s.

*I am indebted to many people who contributed to this article. I particularly would like to thank Takao Gunji, Francis Bond, Dan Flickinger, Melanie Siegel and Timothy Baldwin for a lot of comments and support.

Theoretical and Applied Linguistics at Kobe Shoin 8, 37–51, 2005.

(1) Productive and compositional V_1-V_2s
   a. *aruki-kakeru* (walk-be.about.to) 'be about to walk'
   b. *ai-sobireru* (meet-fail) 'fail to meet'
   c. *yomi-ayamaru* (read-mistake) 'make a mistake in reading'

(2) Less productive and less compositional V_1-V_2s
   a. *odori-tukareru* (dance-get.tired) 'get tired from dancing'
   b. *tobi-okiru* (jump-get.up) 'get up swiftly'
   c. *tataki-waru* (hit-break.in.half) 'break in half by hitting'

(3) Idiosyncratic V_1-V_2s
   a. *kuri-kaesu* (turn.over-give.back) 'repeat'
   b. *uti-kiru* (hit-cut) 'abort'
   c. *tori-midasu* (take-disturb) 'become upset'

The V_1-V_2s listed in (1) are productive, compositional, and transparent as to how their meanings are constructed from their component verbs. Semantically speaking, the V_2s in (1) take V_1's meaning as a semantic argument, or embed V_1's semantics. The V_1-V_2s illustrated in (2) are compositional in some way, but it seems difficult to find a regularity governing all the V_1-V_2s. In (2a), we find that a causation relation holds between *odoru* (dance) and *tukareru* (get.tired), but in (2b), *tobu* (jump) describes the manner in which someone gets up. Besides, these V_1-V_2s are restricted in variation; while we can say *hare-wataru* (clear.up-spread), we would never say something like **kumori-wataru* (cloud.up-spread), even though it makes sense semantically or pragmatically. (3) shows us V_1-V_2s that are non-compositional and highly lexicalized. In the V_1-V_2 in (3b), *uti-kiru* (hit-cut) 'abort', for instance, neither *utu* nor *kiru* contributes their meanings to the compound's meaning 'abort'. V_1-V_2s of this kind are much more restricted in variation than those in (1) and (2).

In spite of their pervasiveness, variety, and complexity, little attention has been paid to V_1-V_2 compounds in previous computational grammars of Japanese (Mitsuishi et al., 1998; Ohtani et al., 2000; Siegel & Bender, 2002; Masuichi & Ôkuma, 2003). Siegel and Bender (2002), for example, merely try to list all V_1-V_2s in the lexicon, identifying them as single words. However, it is certain that this exhaustive listing approach would suffer from undergeneration because of the remarkable productivity of some types of V_1-V_2s.[1] Consider the examples in (4).

(4) a. *tabe-aruku* (eat-walk) 'eat around'
   b. *tabe-aruki-tuzukeru* (eat-walk-continue) 'continue to eat around'
   c. *tabe-aruki-tuzuke-sobireru* (eat-walk-continue-fail) 'fail to continue to eat around'
   d. *tabe-aruki-tuzuke-sobire-hazimeru* (eat-walk-continue-fail-begin) 'begin to fail to continue to eat around'

All V_1-V_2 compounds in (4) are grammatical and really productive, which indicates that the exhaustive listing approach to any kind of V_1-V_2 compounds is not realistic. We must distinguish between productive V_1-V_2s and non-productive V_1-V_2s and provide an account of the

[1] Sag et al. (2002) and Baldwin and Bond (2002) call this problem a **lexical proliferation problem**.

former which captures proper generalizations. On the other hand, dealing with any kind of V_1-V_2 in a fully compositional way without distinction between productive and non-productive, a simple concatenation approach, would face the problem of overgeneration. As mentioned above, not all imaginable combinations of verbs, like **kumori-wataru* (cloud.up-spread), are attested. The simple concatenation approach cannot rule out such impossible cases. Besides, such an approach has no way of predicting differing compositions of meanings of V_1-V_2s. Indeed, the meanings of V_1-V_2s in (1), (2), and (3) seem to be formed by different rules or principles. Especially, the V_1-V_2s in (3) seem idiomatic and not decomposable.[2]

Considering the inadequacy of the simple solution, it is clear that we need a sophisticated linguistic treatment for V_1-V_2 compounds in Japanese.

2. Background: Hashimoto (2004)

2.1 The Analysis of V_1-V_2 Compounds

In order to deal with V_1-V_2 compounds in Japanese properly, in a previous paper (Hashimoto, 2004), I proposed an engineering oriented analysis. In that thesis, I made use of the analyses and observations by Kageyama (1993) and Matsumoto (1996) but arranged them according to four criteria proposed by Hasida (1997), by which we can judge a linguistic theory to be suitable for NLP. As a result, I classified V_1-V_2s into eight categories as follows.

(5) Classification of V_1-V_2 compounds (Hashimoto, 2004)

Syntactic V_1-V_2 compounds
1. A type
2. B type
3. C type

Lexical V_1-V_2 compounds
4. Right headed V_1-V_2
5. Argument mixing V_1-V_2
6. V_1-V_2 with semantically deverbalized V_1
7. V_1-V_2 with semantically deverbalized V_2
8. Non-compositional V_1-V_2

Note first that each of the eight categories belongs to either of the two types: **syntactic V_1-V_2 compounds** or **lexical V_1-V_2 compounds**. This division of V_1-V_2 compounds into two types was first proposed by Kageyama (1993). For a syntactic V_1-V_2 compound, the two component verbs are combined in the syntax, while lexical V_1-V_2 compounds are formed in the lexicon. In sum, the V_1-V_2s in (1), which are fully syntactically productive and semantically compositional, are all syntactic V_1-V_2 compounds. On the other hand, some lexical V_1-V_2 compounds such as those in (2) show productivity and compositionality, but others like those in (3) seem idiomatic.[3] Not only a grammatical theory but also a computational grammar should account for these characteristics of V_1-V_2 compounds with their varying degrees of syntactic productivity and semantic compositionality.

Next, let us look more closely at my analysis of lexical V_1-V_2 compounds.[4] Examples of non-compositional V_1-V_2s are shown in (3). Lexical V_1-V_2s show differences in their productivity and compositionality. Above all, as the name indicates, the non-compositional V_1-V_2s

[2]These problems are called the **overgeneration problem** and the **idiomaticity problem** by Sag et al. (2002) and Baldwin and Bond (2002).

[3]For a detailed discussion of the difference between syntactic V_1-V_2s and lexical V_1-V_2s, see Kageyama (1993).

[4]I will not give the details for syntactic V_1-V_2s in this paper. For the details, see Hashimoto (2004, §3.5).

are totally lexicalized since neither V_1 nor V_2 contributes to the meaning of the V_1-V_2. Thus, in Hashimoto (2004), I treated them as not decomposable, i.e. single words, and entered each of them as a whole into the lexicon. In contrast, the other four types show compositionality in some way or other with differing constraints of composition, and hence I posited compounding rules to deal with them. Most of the rules involve an ARG-ST (ARGument-STructure) proposed by Imaizumi and Gunji (2000) that allows us to distinguish between external arguments and internal arguments, that is to say, between agentive verbs and nonagentive verbs.[5]

First, **Right headed V_1-V_2s** are licensed as long as the two component verbs share arguments that agree in the external / internal distinction. For instance, *tataki-wareru* (hit-be.broken.in.half) 'be broken in half by someone's hitting' is licensed as a Right headed V_1-V_2 since both the V_1, *tataku* (a *monotrans* verb), and the V_2, *wareru* (a *monounac* verb), take an internal argument, which is shared by the two verbs in compounding. On the other hand, the pragmatically plausible V_1-V_2, **hasyagi-wareru* (make.merry-be.broken.in.half) 'be broken in half by someone's making merry' is impossible because the V_1, *hasyagu*, is a *unergative* verb, and thus the two component verbs share no argument.

Next, roughly following Matsumoto (1996), I analyzed **Argument mixing V_1-V_2s** as consisting of a *monotrans* or *ditrans* V_1 and a *monotrans* V_2. In addition, the V_2 must be of a type that expresses spatial motion such as *aruku* (walk) and *mawaru* (go around), while the V_1 must not be.[6] They show a peculiarity in that they can take an object argument from either the V_1 or the V_2. A typical example of an Argument mixing V_1-V_2 is *tabe-aruku* (eat-walk) 'eat around,' where the V_1 is a *monotrans* non-motion verb and the V_2 is a *monotrans* motion verb. The V_1-V_2 can take either an object that expresses something to eat (the case where the V_1 contributes its object argument) or another object that represents a location of moving (the case where the V_2 contributes its object (locative) argument).

The third type of lexical V_1-V_2 compound in (5), **V_1-V_2s with semantically deverbalized V_1**, have been analyzed by Tagashira and Hoff (1986), Kageyama (1993), and Matsumoto (1996). Roughly speaking, they all seem to consider the V_1 of the V_1-V_2 a prefix which attaches to the V_2 and loses its original verbal meaning. Furthermore, as Kageyama (1993) points out, the V_1 emphasizes the content of the V_2. Examples of such V_1s include *kaku* (scratch), *hiku* (pull), and *sasu* (thrust). The V_1-V_2 is different from the other three compositional lexical V_1-V_2s in that the compounding of V_1 and V_2 does not seem to make reference to any ARG-ST information. Therefore, those prefix V_1s can attach to both agentive verbs, as illustrated by *kaki-midasu* (scratch-disturb), and nonagentive verbs (except for *argless* verbs), as shown by *kaki-kumoru* (scratch-cloud.up).

Finally, **V_1-V_2s with semantically deverbalized V_2** have a semantic structure in which the semantics of the V_2 embeds the V_1's semantics (Kageyama, 1993) or the V_2 takes on an adverbial meaning that modifies the V_1 (Matsumoto, 1996). Examples of verbs which can act as V_2 for such compounds are *wataru* (spread) and *konasu* (deal with). In contrast to the V_1-V_2 with semantically deverbalized V_1, the V_1 and V_2 of the V_1-V_2 with semantically deverbalized V_2 must agree in agentivity, as Kageyama (1993) notes. Consequently,

[5]Following Imaizumi and Gunji (2000), Hashimoto (2004) classifies verbs into *argless* (verbs without arguments), *monounac* (mono-unaccusative, i.e. verbs with one internal argument), *diunac* (di-unaccusative, i.e. verbs with two internal arguments), *unergative* (verbs with one external argument), *monotrans* (mono-transitive, i.e. verbs with one external argument and one internal argument), and *ditrans* (ditransitive, i.e. verbs with one external and two internal arguments). Obviously, the first three types constitute nonagentive verbs, while the other three types belong to agentive verbs.

[6]Note that spatial motion verbs can take an accusative object that represents the location through which the motion takes place. Thus, in the framework of Hashimoto (2004), they are considered to be transitive verbs.

though *hibiki-wataru* (ring.out-spread), which consists of the two *monounac* verbs, is possible, *sakebi-wataru* (shout-spread) is impossible because it is formed from an *unergative* V_1 and a *monounac* V_2, resulting in a violation of the agentivity constraint.

Using the LKB system (Copestake, 2002), I implemented my analysis of V_1-V_2 compounds in Japanese in a large-scale computational grammar of Japanese, JACY (Siegel, 1998, 1999, 2000a, 2000b; Siegel & Bender, 2002). Then I conducted an evaluation experiment using the [incr tsdb()] system (Oepen & Carroll, 2000) and the Lexeed corpus (Kanasugi et al., 2002; Kasahara et al., 2004). For the evaluation, I prepared two versions of JACY; one was the original JACY, JACY-plain, without an implementation for V_1-V_2s, but with 1,325 V_1-V_2 entries in the lexicon, and the other was Hashimoto's (2004) version, JACY-vv, which includes an implementation of the V_1-V_2 analysis but from which the V_1-V_2 entries had been removed except for some non-compositional V_1-V_2s. The result showed that JACY-vv has broader coverage and shows less ambiguity than JACY-plain. JACY-vv's broader coverage is surprising since JACY-plain was given as many as 1,325 V_1-V_2 entries in the lexicon. I suspect that this was because of the remarkable productivity of some types of V_1-V_2 compounds; they required the generalization of V_1-V_2 compounding.[7]

In summary, Hashimoto (2004) proposes a theoretically precise and yet broad coverage treatment of V_1-V_2 compounds in Japanese. The treatment can also generate a fine-grained semantic representation of V_1-V_2 compounds, which would help NLP systems to be more precise. Note that deep linguistic analyses and observations brought us these advantages.

2.2 The Overgeneration Problem

Hashimoto (2004) was a successful engineering oriented approach to V_1-V_2 compounds in Japanese. Nevertheless, it faces a problem: the overgeneration of V_1-V_2s with semantically deverbalized V_1. In (6), there are ungrammatical V_1-V_2s with semantically deverbalized V_1 that have a "synonymous" grammatical counterpart.

(6) Unattested "synonymous" V_1-V_2s with semantically deverbalized V_1

a. **hiki-yuzuru* (pull-give)
cf. *hiki-watasu* (pull-give) 'give'

b. **sasi-kimeru* (thrust-decide)
cf. *sasi-sadameru* (thrust-decide) 'decide'

c. **tori-hanasu* (take-let.out)
cf. *tori-nigasu* (take-let.out) 'let something get away'

My analysis cannot rule out these ungrammatical V_1-V_2s since it basically attaches any deverbalized V_1 to any verb.[8]

Furthermore, the compounding rule for V_1-V_2s wrongly construes some Right headed V_1-V_2s as V_1-V_2s with semantically deverbalized V_1 when the V_1 is one of the verbs like *hiku*, *sasu*, and *toru*, which can act as semantically deverbalized V_1. (7) includes examples of such Right headed V_1-V_2s.

(7) Right headed V_1-V_2s that are incorrectly given deverbalized V_1 interpretations

a. *hiki-nuku* (pull-pull.out) 'tear something out of'

[7]The reduction in ambiguity was in some cases the result of the restricted nature of my analysis of syntactic V_1-V_2 compounds. For details, see Hashimoto (2004, chapter 4).

[8]Remember that my analysis of lexical V_1-V_2s relies solely on ARG-ST, and that according to Kageyama (1993) there is no restriction on V_1-V_2s with semantically deverbalized V_1 in terms of ARG-ST.

b. *sasi-korosu* (thrust-kill) 'kill by thrusting'
c. *tori-hazusu* (take-unbolt) 'detach'

In the examples in (7), both the V_1 and the V_2 contribute their original verbal meaning to the V_1-V_2. That is, these V_1s are not deverbalized in spite of their surface form which is identical to one of those semantically deverbalized V_1s. However, not only the compounding rule for the Right headed V_1-V_2 but also that for the V_1-V_2 with semantically deverbalized V_1 applies to them. This happens because the latter rule is triggered only by V_1's surface forms, like *hiki*, *sasi*, and *tori*. Note that the V_1-V_2s in (7) are problematic not only for natural language generation but also for parsing; they create a lot of spurious ambiguities.[9]

In what follows, I will propose an alternative analysis of V_1-V_2s with Semantically Deverbalized V_1 and show its advantages through an evaluation experiment.

3. An Alternative Analysis of V_1-V_2s with Semantically Deverbalized V_1

Hashimoto (2004) analyzed semantically deverbalized V_1s as prefixes that attach to both agentive and nonagentive verbs based on the observation of Kageyama (1993). The observation might be correct from the linguistic point of view. However, if the analysis is used for NLP problems, it suffers from overgeneration as mentioned in the previous section. This is a divergence between linguistics and NLP; linguistics tries to generalize phenomena as much as possible, while NLP prefers robustness with respect to naturally occurring texts or speech even if this means a loss in parsimony.

Before presenting my alternative analysis of V_1-V_2s with semantically deverbalized V_1, we examine their characteristics in more detail. First, since the combination of V_1 and V_2 is not constrained by ARG-ST, you might think that the V_1-V_2s are formed freely and hence are productive. But, in fact, their productivity is quite limited.

(8) a. *hiki-watasu* (pull-give) 'give'
b. **hiki-sadameru* (pull-decide) '?'
c. **hiki-nigasu* (pull-let.out) '?'

(9) a. ?*sasi-watasu* (thrust-give) 'hold forth'
b. *sasi-sadameru* (thrust-decide) 'decide'
c. **sasi-nigasu* (thrust-let.out) '?'

(10) a. **tori-watasu* (take-give) '?'
b. **tori-sadameru* (take-decide) '?'
c. *tori-nigasu* (take-let.out) 'let something get away'

As shown in (8) – (10), we can say *hiki-watasu* (8a), *sasi-sadameru* (9b), and *tori-nigasu* (10c). But grammaticality degrades sharply if we use different semantically deverbalized V_1s. This implies that the V_1 and V_2 of the V_1-V_2 with semantically deverbalized V_1 collocate so tightly that the V_2 does not allow other deverbalized V_1s. In other words, they seem to be highly lexicalized in a way similar to non-compositional V_1-V_2s. Nevertheless, unlike non-compositional V_1-V_2s, they show (partial) semantic compositionality; the V_2 retains the verbal

[9]Hashimoto (2004, p.115) reported that 9 out of 133 V_1-V_2 compounds ($\simeq$ 6.77%) in the subset of the Lexeed corpus were V_1-V_2s with semantically deverbalized V_1.

meaning, and the semantically deverbalized V_1, though it loses the original verbal meaning and became a kind of modifier, emphasizes the V_2's content.

Because of the semantic compositionality, Hashimoto (2004) posited a prefixation rule for the V_1-V_2s. Figure 1 illustrates the analysis. The left side of Figure 1 shows the application of the prefixation rule,[10] while the right side is the semantic representation of the compound.[11] Roughly speaking, the semantic representation says that the V_1, the semantics of which is

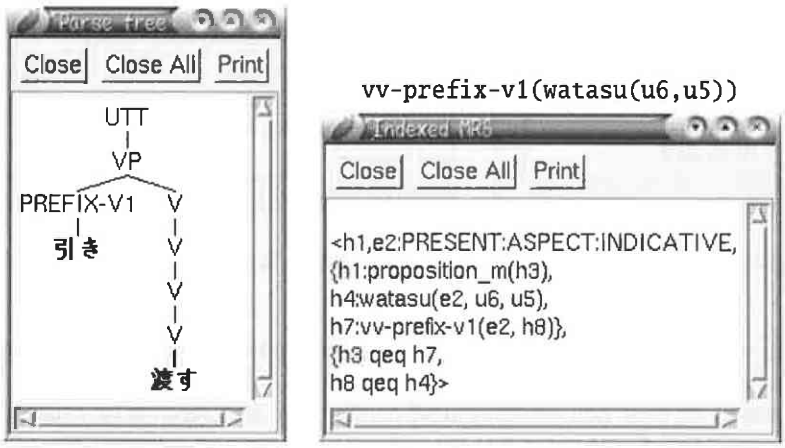

Figure 1: Hashimoto's (2004) analysis of V_1-V_2s with semantically deverbalized V_1

represented by the predicate `vv-prefix-v1`, emphasizes the V_2's content represented by the `watasu` (give) predicate. Clearly, the prefixation rule correctly captures their (partial) semantic compositionality. However, the rule cannot account for the V_1-V_2s' restricted productivity.

My alternative analysis of V_1-V_2s with semantically deverbalized V_1 can deal with both the restricted productivity and the (partial) semantic compositionality. The basic idea is that we do away with the prefixation rule and treat those V_1-V_2s as totally lexicalized, i.e., single words, in the same way as non-compositional V_1-V_2s to cope with their limited productivity and yet give them a compositional semantic representation. Figure 2 illustrates the alternative analysis. Note that, as described in the left side of Figure 2, the V_1-V_2, *hiki-watasu*, is treated as a single word but is given a semantic representation that is almost identical to that illustrated in Figure 1.[12] From this semantic representation, we can correctly learn that *hiki-watasu* (pull-give) basically means *watasu* 'give' with the V_1, *hiki*, semantically deverbalized and emphasizing the V_2's content. This solution might look tedious, but it will turn out to be a better analysis in the evaluation section below.

However, this solution necessarily causes an engineering problem. That is, we would have to enumerate all V_1-V_2s with semantically deverbalized V_1 in the lexicon in order to make the solution feasible when used for NLP problems. Doing this manually would be very time-

[10]Though the binary branching node is labeled VP, the category of the node is, in fact, *word*. But this is irrelevant to the discussion in this paper.

[11]The framework of JACY semantics is based on **Minimal Recursion Semantics** (**MRS**). For details, see Copestake et al. (1999, 2001), Flickinger and Bender (2003).

[12]The new analysis has one technical problem. The main proposition, `h1:proposition_m(h3)`, should have been identified as *vv-prefix-v1(watasu(u6,u5))*, namely `h7:vv-prefix-v1(e2,h4)`, as a whole rather than only *watasu(u6,u5)* that is represented by `h4:watasu(e2,u6,u5)`. In other words, {`h3 qeq h7`} rather than {`h3 qeq h4`}.

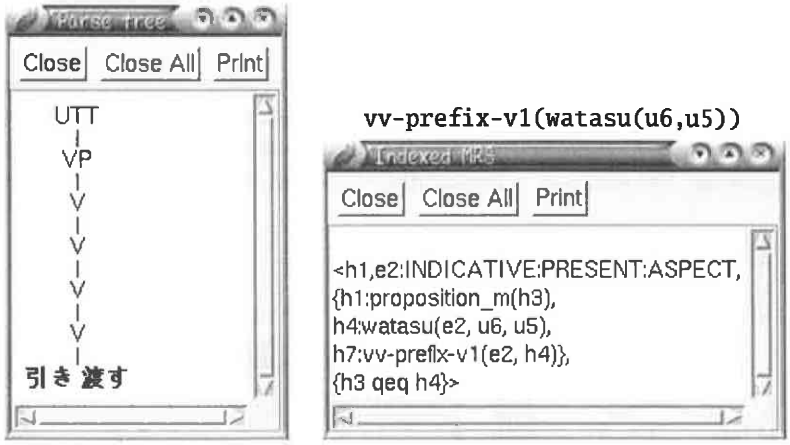

Figure 2: The alternative analysis of V_1-V_2s with semantically deverbalized V_1

consuming, so we need some automatic way of collecting the V_1-V_2s from corpora.

The same is true of non-compositional V_1-V_2s as I discussed in §5.1.2 in Hashimoto (2004); because of their non-compositional nature, they also have to be enumerated in the lexicon. In the thesis, I speculated that the automatic methods of detecting non-compositional English phrasal verbs used by Lin (1999), Bannard et al. (2003), McCarthy et al. (2003) and Baldwin et al. (2003) could be used to help us automatically collect non-compositional V_1-V_2s from corpora. These techniques can be summarized as follows.

(11) Criteria in judging a phrasal verb's compositionality
1. If a phrasal verb is similar to both the head verb and the particle, it is fully compositional.
2. If a phrasal verb is similar to either the head verb or the particle, it is partially compositional.
3. If a phrasal verb is similar to neither the head verb nor the particle, it is non-compositional.

Similarity is measured according to their co-occurrence patterns. In other words, their meanings are approximated in terms of what subjects, objects, and modifiers these verbs take.

Probably we can collect V_1-V_2s with semantically deverbalized V_1 from corpora by means of a similar technique. A criterion for judging whether a V_1-V_2 belongs to the class would be something like this: if a V_1-V_2 is similar to only the V_2, it is a V_1-V_2 with semantically deverbalized V_1. In addition, we can use a further characteristic of these V_1-V_2s, namely the fact that semantically deverbalized V_1s constitute a closed class. As far as I know, there are at most seven verbs that can be semantically deverbalized V_1s. (12) shows examples for each of them.

(12) a. **tori**-*tukurou* (**take**-mend) 'mend'
b. **sasi**-*sadameru* (**thrust**-decide) 'decide'
c. **kaki**-*kumoru* (**scratch**-cloud.up) 'cloud up'
d. **uti**-*nagameru* (**hit**-look.at) 'look at'
e. **osi**-*damaru* (**push**-shut up) 'shut up'

f. **hiki**-*watasu* (**pull**-give) 'give'
g. **tati**-*wakareru* (**stand**-break.up) 'break up'

This characteristic of V_1-V_2s with semantically deverbalized V_1 should be a valuable clue in collecting those V_1-V_2s from corpora. Accordingly, a technique for automatically collecting V_1-V_2s with semantically deverbalized V_1 should make use of the following criteria.

(13) Criteria in judging a V_1 as semantically deverbalized
1. If a V_1-V_2 is similar to only the V_2, the V_1 could be semantically deverbalized.
2. If a V_1 is one of the V_1s in (12), the V_1 could be semantically deverbalized.

Although I have not conducted the experiment of the automatic acquisition of V_1-V_2s with semantically deverbalized V_1 in this paper, I expect that the technique utilizing (13) will show a high accuracy.

3.1 The Applicability of the Alternative Analysis to Other Lexical V_1-V_2s

In Hashimoto (2004), I posited lexical compounding rules for not only V_1-V_2s with semantically deverbalized V_1 but also the other three (partially) compositional lexical V_1-V_2s: Right headed, Argument mixing, and deverbalized V_2 types. However, these three are not equally productive, though they are formed more productively than V_1-V_2s with semantically deverbalized V_1 and non-compositional V_1-V_2s. The degree of productivity comes down in the following order: Right headed V_1-V_2s, Argument mixing V_1-V_2s, and V_1-V_2s with semantically deverbalized V_2.

The question is whether we should give the alternative analysis for each of the three.[13] First of all, Right headed V_1-V_2s are so productive that we can coin compound words of that type so freely as long as they are semantically and pragmatically plausible. (14) includes creative Right headed V_1-V_2s that I discussed in Hashimoto (2004).

(14) a. *hakari-kazoeru* (measure-count) 'measure and count'
b. *osie-mitibiku* (teach-lead) 'lead by teaching'
c. *tuge-siraseru* (report-inform) 'report and inform'

Clearly, it is not a good idea to enumerate all of them in the lexicon. Nevertheless, the alternative analysis can be used to cover some Right headed V_1-V_2s that my analysis incorrectly rules out. In section 2.1, I mentioned that Right headed V_1-V_2s are licensed as long as the two component verbs share arguments that agree in the external / internal distinction. In most cases, this makes a correct prediction, but unfortunately there are a several exceptions to this.

(15) a. *naki-nureru* (cry-get.wet) '(Cheeks) get wet by crying'
b. *ne-midareru* (sleep-jumble) '(Hair) jumbles by sleeping'

Both of the two V_1-V_2s in (15) should belong to Right headed V_1-V_2s but consist of an *unergative* V_1 and a *monounac* V_2. Consequently the two component verbs cannot share arguments and violate the constraint of Right headed type. Note that revising the rule for Right headed type to accept those in (15) necessarily brings about terrible overgeneration; the productivity of those exceptions is very restricted. Obviously, the alternative analysis, which is capable of dealing with compounds that are compositional but not productive, would help; we can enter

[13] In spite of the discussion here, I have not implemented the alternative analysis for the three types. Thus, the evaluation experiment described in the next section deals with only V_1-V_2s with semantically deverbalized V_1.

those exceptions in the lexicon as single words with compositional semantics so that we can extend the coverage and yet avoid overgeneration.

Second, we have found no such exceptions to the Argument mixing type so far. In addition, the productivity is very high.

(16) a. *sakebi-aruku* (shout-walk) 'walk while shouting'
b. *ikari-aruku* (get.angry-walk) 'walk while being angry'

Most Japanese should feel unfamiliar with the V_1-V_2s in (16), but I am sure they will accept them as Japanese compound words. Therefore, I conclude that the alternative analysis does nothing about this type.

Finally, V_1-V_2s with semantically deverbalized V_2 seem to be an intermediate case; positing a compounding rule would lead to overgeneration, but the alternative analysis would suffer from undergeneration. For example, the semantically deverbalized V_2, *-sikiru* (frequently), in (17) does not seem to be productive.

(17) a. *huri-sikiru* (fall-frequently) '(rains) fall heavily'
b. **oti-sikiru* (fall-frequently) '?'
c. ?*nari-sikiru* (ring-frequently) 'ring heavily'

On the other hand, consider the examples in (18)–(20), which indicate that some V_1-V_2s with semantically deverbalized V_2 are moderately productive.

(18) a. *tukai-hatasu* **(use-exhaust) 'use up'**
b. *yomi-hatasu* **(read-exhaust) 'read thoroughly'**
c. **kone-hatasu* **(knead-exhaust) 'knead thoroughly'**

(19) a. *tukai-konasu* (use-master) 'use skillfully'
b. *yomi-konasu* (read-master) 'read skillfully'
c. **kone-konasu* (knead-master) 'knead skillfully'

(20) a. *tukai-mawasu* (use-turn.around) 'use repeatedly'
b. ?*yomi-mawasu* (read-turn.around) 'read repeatedly'
c. *kone-mawasu* (knead-turn.around) 'knead repeatedly'

I suspect that we first have to figure out which semantically deverbalized V_2s are not productive. Then we should posit a compounding rule only for productive ones, and the alternative analysis takes care of those that are not productive. I think that this would be a better treatment for V_1-V_2s with semantically deverbalized V_2, although we need to investigate productivity for each of them.

4. Evaluation

To show the advantages of the alternative approach to V_1-V_2s with semantically deverbalized V_1 over Hashimoto (2004) and the original JACY (Siegel & Bender, 2002), I conducted an evaluation experiment in the same way as in Hashimoto (2004, chapter 4); I investigated the *competence* and *performance* of the three grammars using the Lexeed corpus (Kanasugi et al., 2002; Kasahara et al., 2004) and the [incr tsdb()] system (Oepen & Carroll, 2000). Note that in the grammar profiling context of [incr tsdb()], *competence* means, among other things, **Coverage**, how many sentences the grammar can cover, and **Ambiguity**, the average amount of

structural ambiguity the grammar produces per sentence, and *performance* means how efficiently the grammar works: **Time**, how long the grammar needs to parse one sentence, **Space**, how much memory the grammar consumes to parse one sentence, and **Tasks**, the average number of operations that the grammar conducts to parse one sentence.

First of all, from the subset of the Lexeed corpus, I extracted 219 sentences, each of which contained at least one V_1-V_2 compound. This data was evaluated using each of the three versions of JACY grammar; `The_Original_JACY` (JACY-plain in Hashimoto (2004)), which is not equipped with any V_1-V_2 implementation but has 1,325 V_1-V_2 lexical entries that have been collected manually from several corpora, `Hashimoto_(2004)` (JACY-vv in Hashimoto (2004)), which includes the V_1-V_2 implementation proposed in the thesis but does not contain any V_1-V_2 lexical entries in the lexicon (except for those non-compositional V_1-V_2s), and `The_Alternative`, which is the same as `Hashimoto_(2004)` except for the analysis of V_1-V_2s with semantically deverbalized V_1.[14]

Tables 1 and 2 show the results.[15] Table 1 shows that `The_Alternative` produced less

Table 1: Competence

	`The_Original_JACY`	`Hashimoto_(2004)`	`The_Alternative`
Coverage (%)	52.1	63.5	63.5
Ambiguity (ϕ)	53.41	50.78	46.42

Table 2: Performance

	`The_Original_JACY`	`Hashimoto_(2004)`	`The_Alternative`
Tasks (ϕ)	79,783	137,851	136,281
Time (ϕ)	4.85	6.43	6.34
Space (ϕ)	816,779	995,681	995,232

ambiguity than `Hashimoto_(2004)` in spite of maintaining the coverage of `Hashimoto_(2004)`. As for performance shown in Table 2, `The_Alternative` outperformed `Hashimoto_(2004)` in all the three respects. These results are certainly due to the alternative analysis of V_1-V_2s with semantically deverbalized V_1; getting rid of the overgenerating prefixation rule led to the reduction in ambiguity and the improvement in performance.[16]

5. Conclusion: A Computational Parsimony

Obviously, the most precise way to describe a language is to enumerate possible expressions of the language exhaustively. However, this approach makes linguistics, the science of language, violate the principle of **scientific parsimony**: the principle of explaining a multitude of phenomena by a small number of hypotheses. Therefore, linguistics tries to generalize phenomena as much as possible. However, "accidental" phenomena, without regularity, tend to be ignored.

[14] In this experiment, I assumed that I could collect all the V_1-V_2s with semantically deverbalized V_1 that appeared in the evaluation corpus. Thus, I entered all of them in the lexicon manually in advance.

[15] Here I concentrate on comparing `Hashimoto_(2004)` and `The_Alternative`. For the comparison and the discussion of the difference between `The_Original_JACY` and `Hashimoto_(2004)`, see Hashimoto (2004, chapter 4).

[16] Generally speaking, more rules increase "search space" in which a parser should find a correct analysis. As a result, the parser needs more **Tasks**, more **Time**, and more memory **Space** to perform the job.

Non-compositional V_1-V_2s, for instance, are accidental, and hence no linguist has dealt with them.

In contrast, NLP, the engineering of language, is free from the principle of scientific parsimony. The most important things for NLP are increased precision, broader coverage, and greater efficiency. It is only to achieve these purposes that a notion of parsimony plays a roll in NLP. That is, since enumerating all the possible expressions in a language is impossible, NLP requires a parsimonious description of the language. But we should be aware of the tendency for shotgun generalizations of phenomena to lead to a computational grammar with more ambiguity and worse performance. Thus, a better approach to phenomena which encompass only a handful of expressions will be to list all the expressions in the lexicon or some other component of grammar, even though such an approach looks boring from a linguistic point of view. We may call the trick for grammar development the principle of **computational parsimony**, as opposed to the principle of scientific parsimony.

In this paper, I have provided an alternative analysis of V_1-V_2s with semantically deverbalized V_1 and demonstrated some of its advantages. The approach observes computational parsimony; it exhaustively enumerates all of the V_1-V_2s to cope with their very restricted productivity and semi-lexicalized nature, and yet it successfully accounts for their (partial) semantic compositionality. As a result, the alternative approach attains a reduction in ambiguity and better performance. On the other hand, Hashimoto's (2004) approach, which simply follows Kageyama's (1993) observation that the V_1 can combine with both agentive and nonagentive verbs, violates computational parsimony and overgenerates, resulting in more ambiguity and worse performance.

Appendix: Sample Lexical Entries

In this appendix, I illustrate two lexical entries of V_1-V_2s with semantically deverbalized V_1 that have been implemented in my version of JACY: *hiki-watasu* and *sasi-sadameru*.

As described in section 3., a semantically deverbalized V_1 only emphasizes V_2's content, while the V_2 directly contributes its meaning to the V_1-V_2. Now let's first look at the following entry for the simplex verb *watasu* 'give'.

```
watasu-stem := v1-monotrans-c-non-motion-stem-lex &
 [ORTH ⟨! "渡す" !⟩,
  SYNSEM [LKEYS.KEYREL.PRED 'watasu_rel]].
```

The lexical type for *watasu* is specified as `v1-monotrans-c-non-motion-stem-lex`, and naturally its phonological form, 渡す, and meaning, `watasu_rel`, are stipulated. Next, look at the lexical entry for *hiki-watasu* below. Note that I specified its meaning as basically the same as *watasu*, namely `watasu_rel`, and that I further introduced `vv-prefix-v1-relation` into the semantics, which is meant to emphasize the meaning of *watasu*. The first element of the RELS list, `[ ]`, is a kind of a placeholder for the predicate-argument feature structure of `watasu_rel`. Hence, the meaning of *hiki-watasu* consists of the two predicate-argument feature structures: those for `watasu_rel` and `vv-prefix-v1-relation`.

```
hikiwatasu_prefix-v1-vv := v1-monotrans-c-non-motion-stem-lex &
 [ORTH ⟨! "引き","渡す" !⟩,
  SYNSEM [LKEYS.KEYREL.PRED 'watasu_rel,
          LOCAL.CONT [HOOK [LTOP #lbl,
```

```
                    INDEX #ind],
              RELS ⟨!  [ ], vv-prefix-v1-relation &
                           [ARG0 #ind,
                            ARG1 #lbl] !⟩]]].
```

In the same way, the lexical entry for *sasi-sadameru* is contrasted with that for the simplex verb, *sadameru*.

```
sadameru-stem := v1-monotrans-v-non-motion-stem-lex &
 [ORTH ⟨!  "定める" !⟩,
  SYNSEM [LKEYS.KEYREL.PRED 'sadameru_rel]].


sashisadameru_prefix-v1-vv := v1-monotrans-v-non-motion-stem-lex &
 [ORTH ⟨!  "差し","定める" !⟩,
  SYNSEM [LKEYS.KEYREL.PRED 'sadameru_rel,
          LOCAL.CONT [HOOK [LTOP #lbl,
                            INDEX #ind],
                      RELS ⟨!  [ ], vv-prefix-v1-relation &
                                    [ARG0 #ind,
                                     ARG1 #lbl] !⟩]]].
```

References

Baldwin, Timothy, Bannard, Colin, Tanaka, Takaaki, & Widdows, Dominic (2003). An empirical model of multiword expression decomposability. In *Proceedings of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pp. 89–96 Sapporo, Japan.

Baldwin, Timothy & Bond, Francis (2002). Multiword expressions: some problems for Japanese NLP. In *Proceedings of the Eighth Annual Meeting of the Association of Natural Language Processing, Japan*, pp. 379–382 Japan, Keihanna, Japan.

Bannard, Colin, Baldwin, Timothy, & Lascarides, Alex (2003). A statistical approach to the semantics of verb-particles. In *Proceedings of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pp. 65–72 Sapporo, Japan.

Copestake, Ann (2002). *Implementing Typed Feature Structure Grammars*. CSLI Publications.

Copestake, Ann, Flickinger, Daniel P., & Sag, Ivan A. (1999). Minimal recursion semantics: an introduction. Manuscript, Stanford University: CSLI.

Copestake, Ann, Lascarides, Alex, & Flickinger, Dan (2001). An algebra for semantic construction in constraint-based grammars. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL 2001)*, pp. 132–139 Toulouse, France.

Flickinger, Daniel P. & Bender, Emily M. (2003). Compositional semantics in a multilingual grammar resource. In Bender, Emily M., Flickinger, Daniel P., Fouvry, Frederik, & Siegel, Melanie (Eds.), *Proceedings of the Workshop on Ideas and Strategies for Multilingual Grammar Devel, ESSLLI 2003opment*, pp. 33–42.

Fukushima, Kazuhiko (2003). A neo lexical account for the compounding complexities of v-v compounds in Japanese. draft.

Hashimoto, Chikara (2004). *A Computational Treatment of V-V Compounds in Japanese*. Ph. D. dissertation, Kobe Shoin Graduate School.

Hasida, Kôiti (1997). Information science approach to language. In Ôtsu, Yukio, Gunji, Takao, Takubo, Yukinori, Nagao, Makoto, Hasida, Kôiti, Masuoka, Takashi, & Matsumoto, Yuji (Eds.), *An Introduction to Language Science* (in Japanese), chap. 3. Iwanami.

Himeno, Masako (1999). *The Structure and Semantics of Compond Verbs* (in Japanese). Hitsuji Shobou.

Imaizumi, Shinako & Gunji, Takao (2000). Complex events in lexical compounds. In Itou, Tanake & Yatabe, Shuichi (Eds.), *Lexicon and Syntax* (in Japanese), pp. 33–59. Hitsuji Shobou.

Kageyama, Taro (1993). *Grammar and Word Formation* (in Japanese). Hitsuji Shobou.

Kanasugi, Yuuko, Kasahara, Kaname, Inago, Nozomi, & Amano, Shigeaki (2002). Selection of a basic vocabulary based on word familiarity ratings. In *IEICE Technical Report NLC2002,* No. 27, pp. 21–26.

Kasahara, Kaname, Sato, Hiroshi, Bond, Francis, Tanaka, Takaaki, Fujita, Sanae, Kanasugi, Yuuko, & Amano, Shigeaki (2004). Construction of a Japanese semantic lexicon: Lexeed. In *Information Processing Society of Japan, 2004-NL-159*, pp. 75–82 Tokyo, Japan.

Lin, Dekang (1999). Automatic identification of non-compositional phrases. In *Proceedings of the ACL-1999*, pp. 317–324 College Park, Maryland.

Masuichi, Hiroshi & Ôkuma, Tomoko (2003). Practical analysis of Japanese based on Lexical Functional Grammar (in Japanese). *Journal of NLP*, *10*(2), 79–109.

Matsumoto, Yo (1996). *Complex Predicates in Japanese: A Syntactic and Semantic Study of the Notion 'Word'*. CSLI Publications.

McCarthy, Diana, Keller, Bill, & Carroll, John (2003). Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pp. 73–80 Sapporo, Japan.

Mitsuishi, Yutaka, Torisawa, Kentaro, & ichi Tsujii, Jun (1998). HPSG-style underspecified Japanese grammar with wide coverage. In *COLING-ACL*, pp. 876–880.

Oepen, Stephan & Carroll, John (2000). Performance profiling for grammar engineering. *Natural Language Engineering*, *6*(1), 81–97.

Ohtani, Akira, Miyata, Takashi, & Matsumoto, Yuji (2000). On Japanese grammar based on HPSG—refinement and extension toward computational implementation (in Japanese). *Journal of NLP*, *7*(5), 19–49.

Sag, Ivan A., Baldwin, Timothy, Bond, Francis, Copestake, Ann, & Flickinger, Dan (2002). Multiword expressions: a pain in the neck for NLP. In *Computational Linguistics and Intelligent Text Processing: Third International Conference*, pp. 1–15 Mexico City, Mexico.

Siegel, Melanie (1998). Japanese Particles in an HPSG Grammar. Tech. rep., Verbmobil.

Siegel, Melanie (1999). The syntactic processing of participles in Japanese spoken language. In Wang, Jhing-Fa & Wu, Chung-Hsien (Eds.), *Proceedings of the 13th Pacific Asia Conference on Language, Information and Computation (PACLIC 13), February 10-12* Taipei, Taiwan.

Siegel, Melanie (2000a). HPSG analysis of Japanese. In Wahlster, Wolfgang (Ed.), *Verbmobil. Foundations of Speech-to-Speech Translation* (Artificial Intelligence edition)., pp. 265 – 280. Springer, Berlin, Germany.

Siegel, Melanie (2000b). Japanese honorification in an HPSG framework. In Ikeya, Akira & Kawamori, Masahito (Eds.), *Proceedings of the 14th Pacific Asia Conference on Language, Information and Computation, February 15-17*, pp. 289–300 Tokyo, Japan. Waseda University International Conference Center.

Siegel, Melanie & Bender, Emily M. (2002). Efficient deep processing of Japanese. In *Proceedings of the 3rd Workshop on Asian Language Resources and International Standardization* Taipei, Taiwan.

Tagashira, Yoshiko & Hoff, Jean (1986). *Handbook of Japanese Compound Verbs*. The Hokuseido Press.

Teramura, Hideo (1969). Endings, auxiliary verbs, subsidiary verbs, and aspect—No. 1—. In *Japanese Language and Japanese Culture 1* (in Japanese). Osaka University of Foreign Studies.

Yamamoto, Kiyotaka (1983). *The Structure and Syntax of Compounds: A Study of Japanese Processing for Software Documents* (in Japanese). Information-technology Promotion Agency.

Author's E-mail Address: chashi@sils.shoin.ac.jp
Author's web site: http://sils.shoin.ac.jp/~chashi/