On the Transitivity of Japanese Verbal Doublets: A Model for Stemming and other Applications

Hideo Fujii* and Chisato Kitagawa**

* Department of Computer Science, ** Department of Asian Languages and Literatures
University of Massachusetts, Amherst, U.S.A.

1. Introduction

Stemming is a very popular method for indexing in the system of information retrieval (IR). In English for example, *origins*, *original*, *originate*, and *origination* are stemmed into a common form, *origin*. Thus, when a word is presented as a free term (i.e., not controlled), the linguistic variations by inflections (e.g., plural, tense, etc.) or by derivations (e.g., nominalization by derivational suffixes, etc.) are normalized into a uniquely referable *stem* by cutting off the *endings*. Consequently, only the unique core semantics which is expressed by the stem comes to the substance to be indexed. The stemming operation is typically considered as an empirical procedure such as an ad hoc algorithm (e.g., Porter, 1980) or, a look-up method of the machine readable dictionary (e.g., Krovez, 1993). These approaches are generally not critical enough about the linguistic validity of the boundary between the stem and the ending. The effort of this study is to explore the relation between the stem and the ending, especially we focus on the verb. A morphological classification for Japanese in/transitive[1] verb doublets is proposed. This classification provides a general framework of *level ordering* for verb stems, and it shows a comprehensive applicability for stemming, and other natural language processing. The principle is likely to be applied to many other languages.

In any analytical processing of a sentence, the analysis of the function of the verb always plays a key role in determining the fundamental structure and meaning of the sentence. In syntax, a verb works as a lexical head of the sentence to provide the theta-grid in the argument structure. In other words, the verb governs its arguments cooperating with other functional elements such as auxiliaries, adpositions, etc. There are various linguistic devices that involve this relationship, for instance, in/transitivity of verbs in morphology, causative or passive construction in syntax, and selectivity of theta-roles or cases under semantic constraints. The first stage of sentence analysis is to identify each word and its basic lexical characteristics, so the determination of a verb's

[1] In following text, we use following abbreviations and symbols: in/transitive="transitive and intransitive," Vt="transitive verb," Vi="intransitive verb," ϕ="zero morpheme"; in XXX-yyy, XXX is a stem, yyy is an ending; in XXX-yyy/-zzz, -yyy is a transitive ending, and -zzz is an intransitive one, and *XXX means that XXX is not possible form as ungrammatical.

transitivity and its argument requirements is an essential and important task. This information can be used later in a phase of larger processing. Morphological classification of verbal transitivity which this paper presents gives a theoretical and practical framework.

In Japanese, a considerable number of verb doublets of transitive and intransitive, in which they share a common stem, but with a specific ending pattern for each case. For example, two Japanese words for "break" share the common stem *OR*. The transitive form is *OR-u* and the intransitive form is *OR-eru*. What is the transitive nature of the stem *OR* in these two forms? We tried to answer this question by setting a classification model of these doublets. Our model is very simple (as opposed to the superficial complexity of in/transitive derivations), but the applicability is considerable. There are two kinds of applications: one is analytical, and the other is generative.

Stemming is a straightforward case of analytic application - we extract *OR* from *OR-u* and *OR-eru*. In a character-based indexing algorithm, stemming is approximately done by extracting the Kanji part of the verb. However, when if we adopt a word-based method for more sophisticated language-oriented analysis, we need a well defined method to separate the stem and the ending from a given verb, and to obtain the important in/transitivity information of the form. A problem is that we really did not understand the phenomenon.

On the other hand, the dictionary generation is a generative application of the transitivity classification. It is to convert a single *stem dictionary* which consists of only verb stems into a *verb dictionary* which includes the actual verb entries of transitive or intransitive - we derive *OR-u* and *OR-eru* from *OR*. In a handcrafted verb dictionary, we often find the following two problems: 1) The dictionary often lacks consistency and coverage. By the manual effort, some transitive forms of the intransitive counterparts (or vice versa) are sometimes missing in the dictionary. 2) No information is stored about the relationship between the pair of both forms. If we can write a well defined procedure to produce the both transitive and intransitive forms from the stem, these problems can be naturally solved.

2. Transitive/Intransitive Doublets in Japanese and the Theoretical Difficulties

It is well-known that many Japanese verbs have pair-wise morphological constructions of transitive (V_t) and intransitive (V_i) forms with a shared common stem and its specific verb endings. These ending patterns are derivational in the morphology, but are not inflectional

"...in contrast that .<aru> is especially characteristic for intransitivity, and <asu> is especially characteristic for transitivity, <u> and <eru> take either position depending on the contrastive situation of the in/transitivity. ... We may discover a secret of verb structure in this place." (original in Japanese) In Table 2 (in first three columns), major ending patterns of Japanese V_t-V_i doublets are shown.

Table 2. Ending patterns of Japanese verbs.
[(...) shows frequencies, {...} for selection, and [...] omissible.]

Vt-ending	Vi-ending	Examples	Category
-[a/o] s	-{ϕ/r} (53)	TOB-as/-ϕ (fly), UTU-s/-r (move)	Vi
"	-[r] e (44)	MOY-as/-e (burn),YOGO-s/-re (make/get dirty)	Vi
"	-i (10)	NOB-as/-i (stretch), OT-os/-i (fall)	Vi
-ϕ (5)	-{a/o} r	SAS-ϕ/-ar (stick), TUM-ϕ/-or (stack)	Vt
-e (71)	"	MAG-e/-ar (bend), KOM-e/-or (push/stay)	Vt
-ϕ	-e (25)	YABUR-ϕ/-e (break), NI-ϕ/-e (boil)	Vt
-e	-{ϕ/r} (26)	AK-e/-ϕ (open), TUKAMA-e/-r (catch/hold)	Vi

After Chomsky's Lexicalist Hypothesis (1970), it has been widely acknowledged that a lexicon has an autonomous capability of word formation, in contrast to a simple warehouse of entries as considered in the early model of transformational grammar in 1960s. In other words, the lexicalist model generally recognizes the distinctive operations of word formation in lexicon which are modularly separated from the syntactic ones such as the attachment of inflectional elements. The pro and con of lexicalist approaches are reflecting on the two ways of treatment of Japanese verb doublets. Okitu (1967) called these approaches *dynamic model* and *static model*. The static model treats both transitive and intransitive forms as independent entries in a lexicon. On the other hand, the dynamic model recognizes the derivational process to produce both forms from a common stem. However, the dynamic model is much suffering from a problem of confusing behavior of suffix "-e" which can be simply regarded as neither a transitivizer nor an intransitivizer. On the contrary, the static view misses the relationship between two forms of a doublet, and the lexicon becomes complicated to distinguish every case of either transitive or intransitive with "-e" suffix. But, there is no way to escape the this contradictory problem of "-e," as long as the doublet formation is expressed by choiceless rules on the superficial configuration.

Definitely we need a something more structural mechanism to set up more natural processes in the lexicon. A contribution of this paper is to demonstrate one possible solution of this problem.

In previous studies, there are two typical attempts to solve the "-e" problem. One is Inoue's phonological argument, the other is Okitu's lexical treatment. According to Inoue (1976), Japanese verbal morphology has actually no derivational ending "-e," but a unique intransitivizer "-re," and when "-re" succeeds after a consonant, its "r" sound phonologically disappeared. Therefore, there is no logical contradiction of "-e" phenomenon for her. However, there are counter-examples to her claim such as *Ni-e-ru* (be boiled), *MI-e-ru* (be seen), and *MO-e-ru* (be burned)[4]. Okitu (1967) allowed doublets to be derived from the stems, however he had to introduce a lexical feature to indicate the in/transitivity pattern of every stem. Although he could avoid the "-e" problem, the structure of lexicon became complicated by the pattern indications. Furthermore, he didn't provide a good mechanism for the colloquial morphological productivity which Nishio (1988) pointed out as seen in *SIRAB-ar-u* derived from *SHIRAB-e-ru*. In short, the establishment of dynamic model was not fully successful in previous studies, and this paper demonstrates a possibility of a strong position of dynamic model.

At the end of this section, we note a confusing situation in the analytical task which is caused within both syntax and lexicon. Let us observe following two sentences:

(1) Ie-ga UR-e-ru. (Houses-NOM sell-INTRANS/POTENTIAL-PRESENT)

(2) Ie-ga UR-are-ru (Houses-NOM sell-PASSIVE/POTENTIAL-PRESENT)

Here, (1) (*UR-er*) is a lexical item of intransitive verb, and (2) (*UR-are*) is a syntactic construction of passive, and this first form implies an idiosyncratic extra meaning - "many houses are sold." This is a similar situation with so called *direct* and *indirect causative* which Shibatani's (1972) once argued to defend the Lexicalist Hypothesis.

In the recent development of lexicalist models, the word formations can take place even in elsewhere from lexicon such as in the syntactic component as seen in Kageyama's *Modular Morphology* (1993). However, even in Kageyama's model, the domains of syntactic and lexical manipulations are specifically characterized by each way such as in *verbal compounds* in Japanese. Therefore, the distinction of syntactic and lexical components is a significant property of modular architecture of the language. On contrary, when we analyze and interpret a given sentence in practice, we have to face an interweaving situation of syntax and lexicon in our application system.

---

[4] The corresponding transitive form of this verb is *MO-su*. There is a related but another pair *MOY-asu/-eru*.

For the above examples (1) and (2) of "house is sold," we can ask the same question: "who sell the house?" because both sentences have the same number of valence by intransitivity in (1), and by passive construction of (2).

3. A Stem Classification and the Rule for In/transitive Doublets in Japanese

In general, the verb form in Japanese is not a definitive information to decide the in/transitivity (i.e., +transitive or +intransitive), unless the verb has a suffix of genuine transitivizer (-as) or intransitivizer (-ar/-i)[5] . For example, *SAK*-ϕ (bloom: V_t) and *SAK*-ϕ (tear-off: V_i) have the same phonological value, but different transitivities. As we described in the previous section, there are many Japanese in/transitive verb doublets which share common endings. Although these endings show patterns, they look confusing.

Before we start discussing how to classify these Japanese verb patterns, let us describe the motive to construct our classification in general term. There are three major criteria for an adequate classification system of verb stems, namely: i) *Sufficiency & uniqueness*, ii) *recognizability*, and iii) *generatability.*

Sufficiency and uniqueness require that any verb stem should belong to a class and only one class. Recognizability requires that the combination pattern of possible endings determines the class of the stem. In a better classification, even a partial combination of endings gives good evidence to predict the class. Generatability requires that a class can produce the possible endings for the stem as narrow as possible to the real. If we attach various endings to a stem mechanically, both real and unreal forms in the lexicon may be generated. Complete elimination of this problem may be difficult because of the derivational irregularities, however we should acknowledge this over-productive situation as *possible words* in the derivational morphology[6]. In general, the problem of verb doublets was not often discussed from this point of view. Note that when the application of the classification is analytical such as stemming in information retrieval, there are practically no problems of having unreal possible words because there is no chance to encounter such unreal words in the given text.

---

5 It is notable that this "*s*" sound element is common with "*-suru*" (act) of sahen-verb, and "*{a/o/i}r*" is alike to a verb "*{a/o/i}r-u*" (exist).

6 Although Miyagawa's (1989) blocking model at the paradigm structure explains the filtering process of conflicting elements at a same category, it doesn't address the form itself to be uttered.

In Table 2, there is a significant discrepancy - the transitivizer "-[a/o]s" and the intransitivizer "-{a/o}r" never attach to the same stem[7]. Furthermore, while the ending *"-i"* behaves as an intransitivizer like *-ar*, it is able to become a doublet partner of *-(a)s*. Taking these characteristics, we assume a set of categories, $\{V_i, V_t, \underline{V}_i, \underline{V}_t\}$ for verb stems, and define the following classification rules in Table 3:

Table 3. A scheme of transitivity classification of Japanese verb doublets

(1)	$V_i + \phi = V_i$	$*V_i$ + ar	V_i + [a]s = V_t
(2)	$V_t + \phi = V_t$	V_t + ar = V_i	$*V_t$ + [a]s
(3)	$*\underline{V}_i + \phi$	$*\underline{V}_i$ + ar	$\underline{V}_i$ + [a]s = V_t
(4)	$*\underline{V}_t + \phi$	$\underline{V}_t$ + ar = V_i	$*\underline{V}_t$ + [a]s

There are five basic aspects in this rule set: i) in addition to the level of principal in/transitive categories, i.e., V_t and V_i, we recognize an abstract *underbar level* $\underline{V}$, i.e., realized as $\underline{V}t$ and $\underline{V}i$, which doesn't surface as "immature" to become an actual word, but only as a possible word with a non-zero ending, so that it can resolve the discrepancy of behavior of zero ending (ϕ), i.e., some become actual words with zero, and others do not.; ii) a zero ending does not change any categorical status. Thus, adding zero to $\underline{V}$ level cannot surface an actual word, keeping it immature; iii) the genuine transitivizer *-as* transitivizes only an intransitive stem, i.e., V_i and $\underline{V}_i$. The parallel relation holds for the intransitivizer *-ar*. A vacuous attachment (i.e., V_t or $\underline{V}_t$ plus transitivizer, or V_i or $\underline{V}_i$ plus intransitivizer) produces an ungrammatical word form.

As complementary to above definition of ϕ, *-ar* and *-as*, the verb ending *"-e"* yieldingly determines the transitivity only as a counterpart of already established partner's category by a zero or an in/transitivizer. Because of this irresolute property, the ending *"-e"* has exclusive double-use of either removing the underbar of the category (in contrast to zero which cannot lift the underbar), or switching the genuine in/transitive, apparently. But, it lacks the capability to operate both functions at the same time. In addition, the ending *"-i"* has bilateral characteristics of both *"-e"* (i.e., passive determination of counter category of *-as*[8]) and *-ar* (as an intransitivizer). Above all,

---

7 We couldn't find any exceptions so far. There are superficial false cases such as *mawasu* vs. *mawaru* (rotate), and *hitasu* vs. *hitaru* (soak). These should be delimited with stems of *MAWA-*, *HITA-*, respectively, which are satisfactorily classified as Vi. Evidentially, ending patterns like *-as/-or* or *-os/-ar* do not exist.

8 We found only one exception, "MAZ-er" (Vt: Mix) and "MAZ-ir/-ar" (Vi). But, it still holds the class uniqueness.

our classification consists of rules of promotion and inhibition, symbolically stated as functions:

(1) AS(X)=Xt, but, *AS(Xt)
(2) AR(X)=Xi, but *AR(Xi)
(3) Φ(X)=X, but *Φ(X)
(4) E(Xi)=Xt, E(Xt)=Xi, but E(X)=X
(5) *I(X), but I(Xi)=Xi .

These relations are illustrated in Figure 1. The *"-e"* ending has a function either to remove an underbar, or to switch the transitivity for non-underbar items. This relation is shown in Table 4. For more procedural applications, we can derive more explicitly procedural representation for recognition (A), and for generation (B):

(A) Recognition Rules for Verb Stem Classes

IF X+ϕ found in Lexicon (as actual word),
THEN stem=Category(X+ϕ)
ELSE :
  IF X+ar/ir is presented, THEN stem=Xt
  IF X+as is presented, THEN stem=Xi
  IF X+e is presented, THEN :
    IF X+ar found in Lexicon, THEN stem=Xt
    IF X+as found in Lexicon, THEN stem=Xi .

(B) Generation Rules for Verb Stem Classes

IF stem=Xt, THEN :
  when to make Vi: Generate stem+ar
  when to make Vt:
    IF stem is Underbar level,
      THEN Generate stem+e
    IF stem is NOT Underbar level,
      THEN Generate stem+ϕ
IF stem=Xi, THEN :
  when to make Vt: Generate stem+as
  when to make Vi:
    IF stem is Underbar level,
      THEN Generate stem+e/i
    IF stem is NOT Underbar level,
      THEN Generate stem+ϕ .

Table 4. Operations of "-e" suffix.

E(x)	V	V
V_i	V_i	V_t
V_t	V_t	V_i

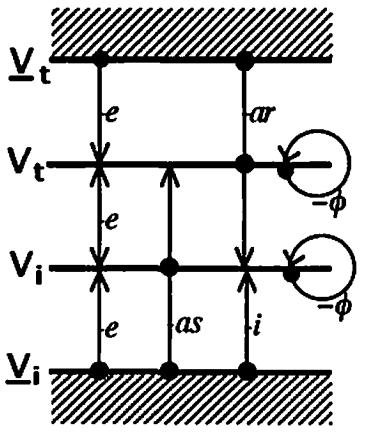

Figure 1. Categorical changes by various doublet suffixes.

V is an important categorical level. V should be an abstract category which is invisible from the syntactic operation. Consequently, the instance of this level cannot exist by itself as a syntactical unit, i.e., word. They are sub-syntactical entities which should be stored in the lexicon. Kageyama (1995) distinguished three levels of Japanese morphological structure: *word*, *stem*, and *root*. In this level order, his stem is a morpheme which may be either bound or free, in contrast to words which are always free, and roots which are always bound. In his system, a simple verb unit like *Tabe(ru)* [=eat] is a root, and their compounds are stems like *Tabe-Aruku* [=hanging-around-to-eat]. Our classification of verb stems is partially compliant with Kageyama's system in that our

verb stems may (with zero) or may not (as $\underline{V}$ level) stand by itself like his stems. However, he did not analyze the inside structure of simple form of a common root and its conjugational endings. The most striking difference of our system from Kageyama's is that our level ordering of V_X and $\underline{V}_X$ does not demonstrate any characterization of compounding. Also, only two levels are required in our verbal classification instead of the three levels in Kageyama's.

On the above bases, can our method indeed classify well all cases of ending patterns of doublets in Japanese? The answer is *yes*, and each class of every ending pattern in Table 2 was imposed at the last column. For the sufficiency condition, stem should belong at least one category because any doublet is either zero, -[a/o]s, or -{a/o}r. For the uniqueness, the exclusive assignment of *-as* and *-ar* endings give strong support of this condition[9]. A stem may have *-[a/o]s* or *-{a/o}r* exclusively, so it cannot belong to both V_t and V_i class at the same time. For recognizability condition, we can determine or narrow the membership of the class by knowing the *-{a/o}r* (or *-[a/o]s*) attachment, and/or zero attachment. For generatability condition, there are several mechanisms to reduce the production of unreal possible words such as the exclusion of zero attachment to an underbar level unit, and the elimination of vacuous ending attachment.

It is our next theme of work to examine the linguistic validity and the computational effectiveness of our proposed rules and procedures of classification.

4. Final Remarks

Our classification described in this paper provided a general framework of abstract level ordering for verb stems to solve the problem of Japanese verb doublets, and it push the model toward a strong position of the "dynamic model." It will give more comprehensive applicability for stemming, and other natural language processing. The range of cross-lingual applicability of our model should also be tested. It is possibly applicable to many other languages such as Korean and other Asian/Oceanian languages, Russian, Finnish, etc. (LINGUIST List, 1996, Jelinek, 1996).

---

9 Indeed, pairs of Vt and Vi do not exist except only a few polysemous doublets such as *hiraku* (open) and *toziru* (close). However, these examples have a third ending form, such as *HIRAK-{ϕ,er}/ϕ* and *TOZ-{i,as}/-i*, respectively. Thus, we can treat the ending of such conflicted pair (e.g., *HIRAK-ϕ* as Vt) as an exception. There are other kinds of conflicts by zeros in some verbal *triplets* such as: *YAM-e/-{ar,ϕ}* (quit) [$\underline{V}$t and Vi]; *YUR-{as,ϕ}/-e* (shake) [Vt and $\underline{V}$i]; *KURUM-{e,ϕ}/-ar* (wrap) [Vt and $\underline{V}$t]. In these cases, we can give higher priority to zero to assign the category as an authentically unmarked form, and others as exceptions. Note that there are false cases such that *KARAM-er/-ϕ*(tangle) vs. *KARAMA-s/-r* , or *YURUM--er/-ϕ*(loosen) vs. *YURUMA-s/-r.*

As an experiment, we are implementing a computer program to automatically generate a verb dictionary from a stem dictionary. The stem dictionary stores the category, Vi, Vt, Vi, or Vt for each stem entry. This program, not only creates the entries of the verb dictionary, but also the argument structure of each entry. It is necessary to handle the problem of *unaccusativity* (Perlmutter, 1978). We will present such problems in other places. We are currently evaluating the performance of the program. Preliminarily, it shows positive results of accuracy and exhaustiveness. At least, the coverage of actual verbs by the automatic generation is superior to popular handcrafted dictionary, and it generates not many unreal possible words.

It is our next theme of work to examine the linguistic validity and the computational effectiveness of our proposed rules and procedures of the classification. Also, it is necessary to investigate the applicability of our framework to the non-doublet verbs.

Acknowledgment

This material is based on work supported in part by the National Science Foundation, Library of Congress and Department of Commerce under cooperative agreement number EEC-9209623.

Bibliography

Bloch, B. (1946), Studies in colloquial Japanese, Part III: Derivation of the inflected words, *Journal of the American Oriental Society*, 66.

Brannen, N. S. (1967), Nihongo-ni-okeru tui-wo nasu ji/ta-doushi to matrkkusu ("Paired in/transitive verbs and matrices in Japanese"), *Kokugo-gaku*, 70.

Chomsky, N. (1970), Remarks on nominalization, R.A. Jacobs and P.S. Rosenbaum (eds.) *Readings in English transformational grammar*. Walthum, MA: Ginn, pp.184-221.

Hayatsu, E. (1989), On the semantic difference between paired and unpaired transitive verbs in Japanese, Gengo Kenkyuu, 95, pp.231-256.

Inoue, K. (1976), *Henkei-bunpou-to nihongo: [Ge] semantic interpretation and other problems ("Transformational grammar and Japanese")*, Tokyo: Taishuukan.

Jacobsen, W. M. (1991), *The transitive structure of events in Japanese*, Tokyo: Kuroshio.

Jelinek, E. (1996), Voice and transitivity as functional projections in Yaqui, In M. Butt and W. Geuder (eds.), The Projection of arguments: Lexical and syntactic constraints, CSLI. (In press).

Kageyama, T. (1993), *Bunpou-to go-keisei ("Grammar and word formation")*, Tokyo: Hitsuji Shobou.

Krovetz, R. (1993), Viewing morphology as an inference process, Proceedings of the 16th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp.191-202). Pittsburgh, PA.

LINGUIST List (1996), Vol-7-1005/1029 (July, 1996).

Miyagawa, S. (1989), *Syntax and semantics: Structure and case marking in Japanese*, San Diego, CA: Academic Press.

Nishio, T. (1954), Doushi-no hasei-ni-tuite ("Verbal derivation: Based on the in/transitive correspondence"), *Kokugo-gaku*, 17.

Okitu, K. (1967), Jidou-ka, Tadou-ka oyobi ryoukyokuka-tenkei ("Intransitivizing, transitivizing, and bipolar derivation: Pairs of in/transitive verbs"), *Kokugo-gaku*, 70.

Perlmutter, D. (1978), Impersonal passives and the Unaccusative Hypothesis, Berkeley Linguistic Society IV, 157-189, University of California.

Porter, M. (1980), An algorithm for suffix stripping, *Program*, 14 (3), pp.130-137.

Sakuma (1967), Gendai-nihongo-no hyougen-to gohou ("Expressions and grammar of modern Japanese"),Tokyo: Kouseisha-Kouseikaku.

Shibatani, M. (1972), Three reasons for not deriving 'kill' from 'cause to die' in Japanese, Syntax and Semantics 1. J. P. Kimball (ed.), Tokyo: Taishukan.

Suga, K. (1980), Heizon-suru jidou-shi/tadou-shi no imi (Semantics of coexisting intransitive/ transitive verbs), *Kokugogaku*, 120.