Patterns of Phonemic Preferences in Japanese non-headed Binary Compounds: What *waa-puro*, *are-kore* and *mecha-kucha* Have in Common

Laurence LABRUNE
(Université Michel de Montaigne—Bordeaux 3)

Key words: word order rules, truncation, dvandva compounds, echo-words, segment force

1. Introduction

In this paper, we examine three types of non-headed Japanese binary compounds: compound abbreviated loanwords such as *poke-mon* (from *poketto monsutaa* "pocket monster"), dvandva, or coordinative compounds, for instance *are-kore* (from *are* "that" and *kore* "this") and ideophonic echo-words such as *mecha-kucha* ("messy"). We analyze their phonemic organization from a statistical point of view, focusing on the segmental organization of each member of the compound, and particularly on the initial segment of each constituent.

Based on a large body of original data, this study provides clear and massive evidence that the phonological nature of the initial segment of a constituent can be crucially relevant in determining the well-formedness of a compound word in Japanese. In the case of echo-words and dvandva compounds, it determines the ordering of the two constituents. In the case of compound abbreviated loanwords (whose order is irreversible), the segmental organization of the base conditions the possibility to derive a shortened form.

The most significant descriptive generalization which can be grasped from the data is that the zero initial (empty onset /'/) is favored at the beginning of an initial constituent and that /k/ is rejected in the same position, while the exact opposite is true at the beginning of a final constituent: /k/ is extremely frequent, while /'/ is nearly absent, all in a statistically significant manner.

This paper is organized as follow: in section 2, we present, discuss and analyze the data concerning compound abbreviated loanwords, which are the starting point of this study. In 2.4. we provide a principled account of the bias observed concerning the relative frequency of segments at the beginning of each constituent, based on the insight that the likelihood for a given segment to occur as an initial is determined by its segmental force, involving phonetic as well as phonological parameters and motivated by morphological and phonological principles. We then proceed to the presentation and discussion of dvandva compounds (3.1) and echo-words (3.2). In 3.1.3. a reply to the analysis provided by Kageyama (1982) concerning the ordering principles of dvandva compounds is given. In section 4, we present evidence for the non-headedness of compound abbreviated loanwords, dvandva compounds and echo-words and discuss a number of other similarities that they share, especially with regard to *rendaku* and compound accentuation. In 5, we offer a conclusion to the paper and provide a review of further issues which require a more thorough investigation.

2. Compound abbreviated loanwords

2.1. General background

Compound abbreviated loanwords (hereafter CALs) are words prosodically derived from a compound base which lexical components are of foreign origin (excluding Chinese). The phonological material retained in the CAL comes from both elements of the base, as can be seen in the examples in (1). CALs therefore consist of two constituents which will be referred to as C1 and C2[1].

1) Transcription of Japanese words follows the Hepburn system, except that long vowels are indicated by doubling the vowel letter. The original form of the

(1)

poke-mon	<	*poketto monsutaa*	{pocket monster}
tere-ka	<	*terehon kaado*	{telephone card}
mo-ga	<	*modan gaaru*	{modern girl}
don-peri	<	*don perinyon*	{Dom Pérignon}
paso-kon	<	*paasonaru konpyuutaa*	{personal computer}
koma-tare	<	*komaasharu tarento*	{commercial talent}

CAL derivation is a frequent and productive process in Japanese. CALs are true Japanese words in the sense that they have been created in Japan and are not used in the source language (except for rare cases of back-borrowing such as *poke-mon*). In many cases, even the base itself is a Japanese lexical creation which does not exist as a compound in the original language. Moreover, the meaning of the compound is not always predictable from the meaning its constituents have in the source language. For instance, the base *komaasharu tarento* {commercial talent} is a "made in Japan" lexical item. It does not refer to a special talent for business, but to a TV actor or singer who specializes in commercials. The semantic, lexical, and phonological processes that CALs undergo in Japanese belong to the central grammar of Japanese, not to the grammar of the source language.

The present research is based on the statistical exploitation of a corpus of 779 CALs collected over the recent years through the exploitation of a wide variety of sources such as dictionaries, advertisements, the press, ordinary conversations and internet sites[2]. To the best of our knowledge, it is the largest CAL corpus ever gathered. Working on a large corpus provides a unique opportunity to draw generalizations and to uncover phenomena which would otherwise remain unnoticed.

Previous works on the morphology and phonology of Japanese abbreviated loanwords have exclusively focused on the analysis of the

base in the source language is given in { }.

2) The corpus can be consulted at http://erssab.u-bordeaux3.fr/labrune.

prosodic aspect of CAL derivation. And indeed, it is a well-established fact that loanword truncation in Japanese is prosodically driven. Following the work by Itô (1990), a number of subsequent studies have demonstrated that phonological well-formedness principles formulated either as minimal word requirements, templates, or prosodic constraints are relevant in order to account for the prosodic structure of loanword truncations. So it is not unexpected to see that about 80% of the CALs belonging to our corpus are overwhelmingly 4-mora long and prosodically structured as a 2+2 morae organization. The correct generalization is that CALs are made of 2 feet, and that each of these feet is binary in morae.

The general pattern for deriving a CAL from a given source word follows the derivation mechanism illustrated in (2), where *poke-mon* is derived from *poketto monsutaa* {pocket monster}. This schema illustrates the process whereby the two leftmost morae of each lexical constituent in the base are extracted in order to derive a four-mora long abbreviation.

(2) CAL derivation: general principles
poketto monsutaa → *poke-mon* {pocket monster}

CAL[$_{C1}$[	po	ke]			$_{C2}$[	mo	n]]			
	|	|				|	|			
	μ_1	μ_2	μ_3	μ_4		μ_1	μ_2	μ_3	μ_4	μ_5
	|	|	|	|		|	|	|	|	|
Base[$_{C1}$[	po	ke	t	to]	$_{C2}$[	mo	n	su	ta	a]]

CALs derived according to the model in (2), that is, CALs which are two-feet long with each foot being bimoraic, left-anchored, and not skipping any morae, amount to about 65% in the corpus.

However, even though prosodic well-formedness constraints play a major role in the derivation of loanword abbreviations in Japanese, the crucial point is that prosody alone cannot account for all the aspects of CAL derivation. As we shall see in the next section, a number of puzzling anomalies pertaining to the segmental build-up of CALs exist. In particular, certain segments display unexpected occurrence rates at the beginning of CAL constituents, constituting evidence that aside from prosodic principles, melodic ones also crucially operate in loanword truncative derivation.

2.2. Method and data

The phonemic pattern of the CALs was investigated in order to calculate the rate of occurrence of the various segments at the beginning of each constituent. The results were then compared to the actual rate of occurrence of each segment at the beginning of the general lexicon of foreign loans in Japanese (hereafter referred to as the "corpus of reference[3]"), allowing us to evaluate the significance of the figures, since all segments do not have the same frequency in the lexicon. For instance, knowing that 10.7% of the words occurring in the corpus of reference begin with /k/, we expect to find about the same proportion of /k/ at the beginning of each CAL constituent. The crucial point is that this expectation is not fulfilled: some initials display a statistically significant anomalous frequency at the beginning of C1 or C2. The results appear in Table 1 and Figure 1 below.

Table 1 compares the observed frequency of each segment in the

3) The corpus of reference is based on all the entries of the *Dictionary of Loanword Usage Katakana—English*, Motvani (1991). It is assumed that any word appearing as an entry in this dictionary can be used as a constituent in a compound, and that any compound thereby obtained is potentially the target of CAL derivation. For instance, the word *kaa* {car}, can appear as a C1 in the CAL base, as in *kaa nabigeetaa* {car navigator}, or as a C2, as in *patororu kaa* {patrol car}. It is of course expected that certain lexemes of the foreign loan lexicon have a higher probability to occur as first member or second member of a compound. For instance, prefixes such as *non* {non} or *suupaa* {super} or adjectives such as *rongu* {long} are more likely to appear as an initial member. However, this has no significant consequence on the overall result of this study because of the size of the corpus, and because there exists no correlation in Indo-European languages between the initial segment of a word and its grammatical or semantic category. Moreover, in the exceptional case where over-representation of a particular lexeme can be considered relevant in explaining the figures, it can be easily detected by examination of the corpus, and will be mentioned as such.

initial position of three distinct sets of units: i) the 3009 words belonging to the corpus of reference; ii) the 779 forms occurring as a C1 in the CAL corpus; iii) the 779 forms occurring as a C2 in the CAL corpus. The initials include all the possible segments appearing as a constituent initial, including the empty onset or “zero consonant” represented as /’/. In order to render the comparison more straightforward, the table provides the percentage of occurrence for each initial segment in each set, as well as the estimated, or expected frequency for each initial. For instance, looking at Table 1 reveals that words beginning with /p/ amount to 8.2% of the corpus of reference. This corresponds to an expected frequency of 64 C1’s and 64 C2’s (8.2% of 779) in the CAL corpus. If we check the actual number of CAL constituents beginning with /p/, we find that the number

Table 1. Frequency of initial segments in CAL constituents, in comparison to expected frequency.

	Corpus of Reference		Corpus of CALs: Constituent 1				Constituent 2			
	Observed		Exp.	Obs.		Deviation Rate	Exp.	Obs.		Deviation Rate
	n	%	*n*	*n*	%		*n*	*n*	%	
p	248	8.2	64	74	9.5	+9.81	64	110	14.1	+45.81
b	240	8.0	62	55	7.1	–7.32	62	72	9.3	+9.68
f	92	3.1	24	21	2.7	–2.84	24	13	1.7	–10.84
m	205	6.8	53	62	8.0	+8.95	53	62	8.0	+8.95
w	43	1.4	11	11	1.4	–0.06	11	3	0.4	–8.06
t	505	6.8	53	51	6.5	–2.05	53	62	8.0	+8.95
d	151	5.0	39	28	3.6	–11.11	39	22	2.8	–17.11
s	406	13.5	105	102	13.1	–3.09	105	87	11.2	–18.09
z	87	2.9	23	23	3.0	+0.49	23	20	2.6	–2.51
n	76	2.5	20	42	5.4	+22.29	20	7	0.9	–12.71
r	259	8.6	67	64	8.2	–3.07	67	61	7.8	–6.07
k	323	10.7	84	57	7.3	–26.59	84	171	22.0	+87.41
g	101	3.4	26	24	3.1	–2.17	26	27	3.5	+0.83
y	30	1.0	8	9	1.2	+1.29	8	2	0.3	–5.71
h	141	4.7	37	32	4.1	–4.54	37	37	4.8	+0.46
’	402	13.4	104	124	15.9	+19.93	104	23	3.0	–81.07
Total	3009	100	779	779	100	0	779	779	100	0
Standard Deviation						*11.21*				*33.17*

of C1's starting with /p/ amounts to 74 (9.5% of the C1 set), whereas the number of C2's starting with /p/ amounts to 110 (14.1% of the C2 set).

Anomalous frequency rates for a given constituent initial are detected by comparing its absolute deviation rate (dr) with the standard deviation which appears at the bottom of Table 1. The absolute deviation of each initial is obtained by subtracting the estimated frequency from the observed frequency. Accordingly, only those initials with a deviation rate higher in absolute value than the standard deviation rate are anomalous in a statistically significant way. Figure 1 provides a histogram representation of the data in Table 1.

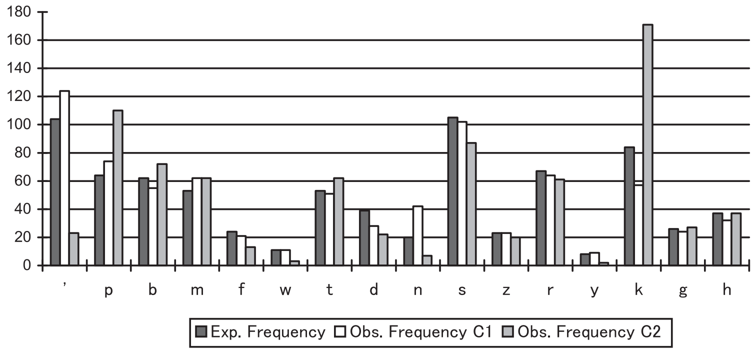

Figure 1. Observed frequency of initial segments in CAL constituents, in comparison to expected frequency (in absolute value).

In Figure 2 below, we also provide a ranking of all the C1 and C2 initials according to their deviation. Only the segments at both ends of the scale, in boldface, have a deviation rate statistically significantly higher than the standard deviation rate. The rankings in Figure 2 should be compared with that of the initials in the corpus of reference (Figure 3).

a. C1 initials:

–															+
k	d	b	h	s	r	f	g	t	w	z	y	m	p	’	**n**

b. C2 initials:

–															+
’	s	d	n	f	w	r	y	z	h	g	m	b	t	**p**	**k**

Figure 2. Classification of CAL C1 and C2 initials, from less to more frequent, based on standard deviation (in bold: initials with a statistically significant anomalous frequency).

–															+
y	w	n	z	f	g	h	d	m	t	b	p	r	k	’	s

Figure 3. Classification of initials in the corpus of reference, from less to more frequent, based on absolute frequency.

2.3. Comments and discussion: the asymmetries between C1 and C2 initials

A first general observation to be made about these data concerns the significant difference between C1 and C2. C1 presents a standard deviation rate of 11.21 whereas C2 has a much higher one: 33.17. This signifies that the phonemic disparities are more blatant at the beginning of C2. The results of the χ^2 test confirm this result (C1: χ^2=46.125, p<0.001, df=15; C2: χ^2=226.475, p<0.001, df=15), showing that the distribution and frequency of the initial in CALs is not in accordance with what is expected under the null hypothesis.

Regarding C1, observe that segments with a deviation rate superior to 11.21 in absolute value are /n/ (dr=+22.29) and /’/ (dr=+19.93) on one hand, and /k/ (dr=–26.59) on the other hand. The actual number of C1’s beginning with /n/ is 42 (5.4%) in contrast with the expected 20 (2.5%), and that of /’/ is 124 (15.9%), to be compared with the expected 104 (13.4%). We therefore conclude that /n/ and /’/ are significantly more frequent than expected. On the contrary, /k/ occurs significantly less than expected. /k/-beginning C1’s amount to 57 (7.3%) in contrast with the expected 84 (10.7%). At the beginning of C1, /k/ therefore appears as the most conspicuous initial with regard to the discrepancy between the observed and the expected figures.

Among the three segments /n/, /'/ and /k/ just cited, the case of /n/ deserves further comments. The higher frequency of /n/ as a C1 initial is due to the frequent use of *noo* and *non* (the English negations {no} and {non}) as initial components in loanwords compounds (examples: *noo suri* < *noo suriibu* {no sleeve}, *noo-hazu* < *noo hazubando* {no husband}, *non pori* < *non poritikaru* {non political}, etc.). However, note that no such lexical reason can be invoked in order to explain the observed rates for /'/ and /k/ at the beginning of C1, which means that the case of /n/ should be treated apart from that of the other initials.

The standard deviation rate of C2 equals 33.17, revealing, as already stated, that disparities between the expected and the observed are more salient in C2 than in C1. Segments with a deviation higher than 33.17 in absolute value are /k/, /p/ and /'/, with respective deviations of +87.41, +45.81, and –81.07. At the beginning of C2 there are 171 instances (22%) of /k/, which is much more than the expected figure of 84 (10.7%), and 110 instances of /p/ (14.1%) in comparison with 64 (8.2%) expected. On the contrary, we find only 23 C2's (3%) with initial /'/, when we expect to have 104 (13.4%) instances of it. We therefore conclude that /k/ and /p/ occur far more frequently than expected at the beginning of C2, while /'/ is remarkably less frequent than expected in the same position.

On the basis of the facts considered thus far, it is clear that a number of segments display statistically anomalous frequency rates at the beginning of C1 or C2, and that the most conspicuous discrepancies occur at the beginning of C2. Summing up the statistically significant results so far observed, we can say that /'/ is clearly disfavored at the beginning of C2, whereas /p/ and especially /k/ are highly favored. Note that at the beginning of C2, /k/ is more than twice as frequent as expected, whereas the zero onset /'/ occurs between four and five times less than expected. At the beginning of C1, we note the opposite tendency for /'/ and /k/: /'/ occurs more than expected, whereas /k/ is less frequent than expected.

As for /n/, as already stated, it occurs more than expected, but for lexical reasons.

To conclude, the above data provide massive evidence that, in addition to prosodic ones, phonemic principles of organization exist, which determine the well-formedness of CALs.

2.4. Accounting for the segmental patterns of CAL constituents initials: the Segment Force Hierarchy

In this section, we attempt to provide an explanation for the segment frequency bias observed in the CAL corpus. The anomalies uncovered above are due, we argue, to phonetic and phonological causes: the idea is that the presence of certain phonetic or phonological features in a given segment contributes to its qualification or disqualification as a preferable C1 or C2 initial.

Segmental force. Our claim is that the segmental preferences at the beginning of C1 and C2 are to be interpreted alongside a Segment Force Hierarchy (hereafter SFH).

Note that /k/ and /p/, the two initials which are statistically over represented at the beginning of C2, are [+consonantal], [-continuant], [-voice] and [-coronal], a set of features which characterizes them as the strongest of all segments. The hierarchy under discussion here is defined as some derived phonological property, resulting from a combination of perceptual salience and phonological markedness. A strong segment can be defined as less sonorous, and it is articulated with greater energy, greater tension and greater articulatory force (see Fougeron 1999 for a review of the different proposals concerning segment force). It must also be structurally marked in the phonological sense of the term: in this respect, the velar and labial places of articulation are "stronger" than coronal. Perceptual salience thus explains why voiceless plosives are "strong", while phonological markedness accounts for the fact that dorsal and labial places of articulation are favored over coronal (see Paradis & Prunet, 1991, Prince & Smolensky, 1993, etc. for arguments concerning the unmarked status of coronal)[4], yielding /k/ and /p/, the two voiceless non-coronal plosives of the Japanese system, to stand as the strongest of all consonants. At the other end of the SFH lies the zero consonant /'/, the weakest of all.

Our data do not provide statistically significant evidence as to the hierarchy of particular segments other than /'/, /k/ and /p/, but note that in Figure 2.b, the four consonants which follow /k/ and /p/ are /t/, /b/, /m/ and /g/, that is, segments which are all [-continuant] and are either [-voiced] or [-coronal], a result which conforms to our hypothesis concerning the SFH principles. Due to space limitations, we cannot present and discuss all the problems and implications of the SFH. The SFH should be seen as a theoretical extrapolation based on the data provided by the analysis of the CAL corpus, but further research is needed on the precise characterization and definition of "phonological force", and the position of segments other than /k/, /p/ and /'/ in the hierarchy (see also Murata 1984 for another proposal concerning the "sound hierarchy" in Japanese). However, we shall insist upon the fact that the SFH discussed here is not a reverse version of the sonority hierarchy[5]. The main difference is that the place of articulation is arguably relevant in the SFH whereas it is not in the sonority hierarchy. For instance, we assume that it is because they are [+coronal] and [+voiced] that /d/ and /n/, although [-continuant], are attributed a lower rank than /t/ and /m/ in the SFH, while they classify with the other segments of the same manner of articulation in most versions of the sonority hierarchy. In our SFH, only one coronal is assumed to rank relatively high: /t/. This is because /t/, although [+coronal], is

4) Our data provide evidence that dorsal is more marked than labial, since /k/ is by far the favorite C2 initial, far ahead of /p/, unlike what is assumed in Prince & Smolensky (1993)'s Place Markedness Hierarchy: *Dorsal, *Labial >> *Coronal, where the two constraints *dorsal and *labial are left unranked above *Coronal.

5) Although no general agreement among scholars exists concerning the exact classification of segments in the sonority hierarchy, most authors adopt the following ranking: plosives–fricatives–nasals–liquids–glides–vowels (see Hooper, 1976, Selkirk, 1984, etc.).

[-voiced] and [-continuant], two features which compensate its coronality to a certain degree.

In sum, segmental preferences in CAL initials can be considered to reflect a preference for a strong segment such as /k/ or /p/ at the beginning of C2, and for a weak one, like /'/ (and /n/[6)]) at the beginning of C1, while conversely, strong segments are avoided at the beginning of C1 and weak ones at the beginning of C2.

It is significant to observe that 13 out of the 24 irregular CALs whose C2 does not start with the initial portion of the second member of the base (in other words, CALs whose C2's left margin is not aligned with the base second member's left margin) would have started with a vowel if they had been derived according to the regular pattern. This class of exceptions includes CALs such as *ana-doru* < *anaunsaa aidoru* (**ana-ai*, {announcer idol}) or *sebu-re* < *sebun irebun* (**sebu-ire,* {seven eleven})[7)]. This confirms the existence of a strong ban against vowel-initial C2's. In these examples, a non initial portion of the second member of the base has been used as the CAL C2 in order to avoid a vowel initial C2[8)]. Note, though, that this strategy remains marginal and that it cannot account for

6) The high frequency of /n/ at the beginning of C1 is due to lexical reasons, but note that the preference for /n/ as a C1 initial also conforms to the SFH since /n/, as a voiced coronal, ranks as a segment with lower consonantal force. Interestingly, we shall see later in this paper that /n/ is also favored at the beginning of dvandva compounds and echo-words, even though no lexical factors can be detected for these two groups of words.

7) The 13 examples are: *ana-doru* < *anaunsaa aidoru* {announcer idol}, *baibu-raamu* < *baibureetaa araamu* {vibrator alarm}, *bara-doru* < *baraeti aidoru* {variety idol}, *bide-kon* < *bideo aikon* {video icon}, *goo-hee* < *goo-aheddo* {go ahead}, *mama-doru* < *mama aidoru* {mama idol}, *ona-doru* < *onanii aidoru* {onanie idol}, *paso-doru* < *pasokon aidoru* {pasokon idol}, *sebu-re* < *sebun irebun* {seven eleven}, *chai-doru* < *chainiizu aidoru* {Chinese idol}, *chai-doru* < *chairudo aidoru* {child idol}.

One should also mention *ban-k̲ame* {Bank of America}, in which the final consonant of the first constituent (notice that it is a /k/) is parsed as the C2 initial, presumably to avoid a /'/-starting C2.

8) An anonymous reviewer has suggested that examples such as *ana-doru* should rather be handled as blends. This is a possibility, of course, but in our view, such

the general underrepresentation of /'/ as a C2 initial in the corpus[9].

Hiatus avoidance might also be invoked here as a partial explanation of the patterns observed above: vowel-beginning C2's are avoided because they would cause a succession of two vowels word-internally. However, hiatus avoidance alone does not account for the preference for /'/ at the beginning of C1, nor does it account for the overrepresentation of /k/ and /p/ at the beginning of C2.

Furthermore, it is necessary to verify whether the phonemic patterns uncovered in CALs are not just a reflection of phonotactic preferences occurring at the level of the unabbreviated compound words acting as bases themselves. That is, one should consider the possibility that, say, /k/ is overrepresented in C2 simply because it is overrepresented at the beginning of the second members of the bases which constitute the input for the truncation. In order to test this hypothesis, we verified the proportion of each second constituent initial in binary non abbreviated compounds of Western origin occurring as entries in the *Sanseido Katakana-go Jiten*, 2nd edition, 2000, and which are labeled as 日, i.e. compounds which were coined in Japan and do not exist as such in the source language. Such words include, for example, *aato furawaa* {art flower}, *aisu keeki* {ice cake} or *angureezu soosu* {anglaise sauce}. The investigation was conducted for the /'/ entry and the /k/~/g/ entry (/k/ and /g/ are treated under the same

---

formations are best analyzed as the result of /'/ avoidance as a C2 initial. Actually, we would not be too surprised to find out that blends exhibit the same types of segmental preferences as those which have been found in CALs. This would mean that the distinction between blends of Western origin and CALs might be irrelevant, especially in the case of quadrimoraic forms such as *anadoru*.

For details concerning the principles and criteria for the selection of the items to be included in the corpus, see Labrune (2002).

9) In addition, noteworthy is the fact that /'/-initial forms also occur significantly more than expected in simplex abbreviated loanwords. Simplex abbreviated loanwords beginning with /'/, such as *anime* < *animeeshon* {animation}, *irasuto* < *irasutoreeshon* {illustration}, *ana* < *anaunsaa* {anouncer}, amount to 20.7%, when only 13.4% are expected (data from our Simplex Abbreviated Loanwords corpus).

entry since the *Sanseido Katakana-go Jiten* follows the *kana* order) and compared with the expected frequency figures provided in Table 1. The results appear in Figure 4.

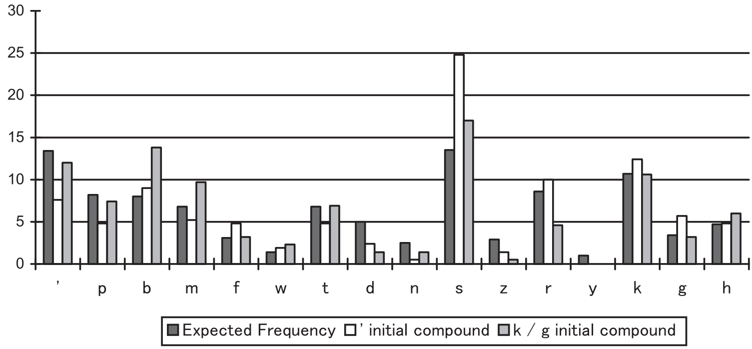

Figure 4. Observed frequency of second constituent initial segments in unabbreviated compound words of Western origin coined in Japan, and starting with /’/ and /k/~/g/ (in %).

The results above show that the phonemic preferences in non-abbreviated compound words are not at all similar to those occurring in CALs. The coronal fricative /s/ is the most frequent of all C2 initials both in /’/- and /k/~/g/-compounds, whereas /k/ represents 12.4% of the C2 initials in /’/-starting unabbreviated compounds, and 10.6% of them in /k/ or /g/-starting ones, in accordance with the expected frequencies. As for /’/, it occurs in respectively 7.6% and 12% of all second members. Noteworthy is the fact that the frequency of /’/ and /k/ are not at all anomalous and do not vary significantly between the first and second member of the compound. So the initials patterns in compound words are clearly different from those occurring in CALs. Therefore, we can confidently conclude that the phonemic initial patterns which emerge in CALs are not a reflection of segmental preferences enforced at the base level.

Why have a strong segment at the beginning of C2? One can assume that for perceptual reasons, salient segments such as /k/ and /p/ stand as optimal boundary cues between the two components of the CAL because of their greater perceptibility and distinctness when they occur between two vowels. This type of organization facilitates the processing of information (Cooper & Ross, 1975, Cutler & Cooper, 1978). C2 initials act not only as lexical and morphological boundaries but also as phonological boundaries. They mark the beginning of a lexeme, the beginning of a compound constituent, and the beginning of a foot. Thus, the presence of a strong consonant at the lexical, morphological and phonological boundary fulfills a demarcative function, facilitating recoverability and accessibility to the lexical, morphological and phonological structure of a given CAL and helping to establish the right correspondence with the source word.

The necessity to enhance the perceptibility and distinctness of the second constituent initial is grounded in the fact that although CALs are morphologically compound lexemes, there is no direct evidence in their phonological structure that they are compound rather than simplex words. First, CALs do not exhibit the accent pattern typical of compound nouns in Japanese: they are unaccented or bear an accent on the initial mora, whereas Japanese compound nouns are generally accented on the initial mora of the second member, or on the final mora of the first member. Second, CALs do not exhibit *rendaku*, whereas a majority of native Japanese compound nouns with a second member starting with /k/, /s/, /t/ and /h/ do. Third, being in majority four-mora long, CALs are relatively short. And quadrimoraicity does not constitute, in Japanese, a length typical of a compound word because many quadrimoraic nouns are simplex words, whereas nearly all nouns which are over five morae long are compounds. Finally, since most CAL constituents do not exist as independent words in Japanese, lexical identification and retrieval are slower than if the correspondence between a CAL constituent and a lexical form were perfect. Take the CAL constituent *kon* for instance, which can stand for the truncated version of *konpyuutaa* {computer}, *konpurekkusu* {complex}, *kontorooru* {control}, and so on. Since *kon* does not occur as a free form in Japanese, its identification as an existing morpheme in *paso-kon* < *paa-sonaru konpyuutaa* {personal computer}, *faza-kon* < *fazaa konpurekkusu* {father complex}, and *rimo-kon* < *rimooto kontorooru* {remote control} takes longer than if it did, when heard for the first time.

Because CALs do not exhibit any formal clue that they are compounds, the presence of a perceptually and phonologically salient segment at the beginning of the second member enhances their identification as lexically compound words in the absence of any other prosodic or lexical clue.

Conversely, the preference for a weak initial at the beginning of a CAL C1 permits maximal prominence contrast between the two CAL constituents, thus warranting a culminative function to the strong C2 initial.

The Japanese facts are somewhat reminiscent of what occurs in Hebrew blends, as described by Bat-El (1996), where the order of two elements in the blend is determined by phonological principles. Bat-El posits the constraint σContSlope that holds that the onset of the syllable in the second element in blends should not be more sonorous than the preceding segment ("the greater the slope in sonority between the onset and the last segment in the immediately preceding syllable the better"). A similar formal approach could be adopted here in order to account for the tendency to prefer a strong segment at the beginning of CALs' C2s, but this treatment would say nothing about the preference for consonantally weak segments at the beginning of C1. Moreover, contrary to Hebrew blends, the order of Japanese CAL constituents is determined by syntactic and semantic factors, and cannot be modified through the derivation, an issue to which we will come back to later on in section 4.

Co-occurrence restrictions. We checked whether dissimilation could be a possible cause to the segment patterning observed in the data. One can conceive that, in order to prevent the co-occurrence of segments sharing similar phonological characteristics at the beginning of C1 and C2, the best strategy would be to reduce the possibility of occurrence of a given initial to a specific position, either C1 or C2. However, we found no statistical evidence to support this supposition. No significant co-occurrence restrictions among natural classes of segments such as manner of articulation and place of articulation were detected. While it is true that there exists a predetermination for certain segments to occur as a preferable C2 or C1 initial, such predetermination does not result from a strategy to avoid certain phonemic combinations at the beginning of C1 and C2. We therefore conclude that there is no link between the initial preferences at the beginning of C1 and those at the beginning of C2, outside of the culminative function.

3. A comparison with dvandva compounds and echo-words

Interestingly, segmental preferences similar to the ones uncovered in CALs can be found in two other morpho-lexical classes of Japanese: dvandva compounds and ideophonic echo-words.

3.1. Dvandva compounds

Dvandva compounds, for instance *shiro-kuro* "white and black", *are-kore* "that and this" or *nishi-higashi* "West and East" consist of a composition of equipollent terms in which both members have equal morphological and syntactic status. The order of dvandva compounds is generally assumed to be conditioned by semantic or pragmatic factors (Kageyama, 1982), but as we shall see, phonological principles play a far from trivial role in determining the relative order of the constituents.

3.1.1. Method and data

We examined the phonological structure of a list of 201 binary dvandva compounds of native or mixed origin (i.e. at least one of the two members of the compound belongs to the Yamato stratum) in order to examine whether the order of the constituents could be due to the phonological nature of the initial segments of the two lexical words occurring as constituents. The results of this investigation appear in Table 2. They amply confirm the existence of significant differences between C1 and C2 with respect to the segments which occur most frequently as constituent initials.

The methodology adopted here is similar to the one followed for CALs. We counted all the initials occurring at the beginning of the first and second constituents of each dvandva compound, and compared the frequencies thus obtained with that of the corpus of reference. The entries of the *Jidai-betsu Kokugo Daijiten Jôdai-hen* (hereafter JKDJ), a dictionary of Old Japanese which includes only lexemes belonging to the native stratum of the language (Yamato stratum) as well as a few ancient Chinese borrowings, were retained as the corpus of reference[10]. Expected frequencies are calculated on the basis of the percentage of each entry initial in the JKDJ.

A few remarks are in order concerning the phonemic characteristics of Yamato Japanese, the stratum to which nearly all our dvandva compounds belong to: /f/ does not exist in this stratum; moreover, /p/ does not occur word-initially and is extremely rare in other positions in modern Yamato words, which is why it does not appear in Table 2. Note that archaic Japanese did possess a /p/, but this sound evolved into the bilabial fricative /ɸ/ and then into /h/ word-initially. Even if the date of this change is subject to controversy (see Hashimoto, 1928, Hamada, 1954, Hayashi, 1992 for different proposals), scholars generally agree that it occurred no later than the 10th century. So, in Archaic and Old Japanese, /h/ and /p/

---

10) Although the JKDJ is an Old Japanese dictionary whereas our list of dvandva compounds belongs to the modern lexicon, we assume that the data provided by the JKDJ were closest to ours in terms of the stratum and grammatical categories they belong to, and that they were therefore the most reliable. First, the constituents of the dvandva compounds are overwhelmingly of Yamato origin (our dvandva corpus contains only ten Sino-Japanese constituents which represent 2.5% of the total), like the lexemes appearing as entries of the JKDJ. Second, the phonemic system of word initials has not changed significantly since the 8th century, except for a few minor transformations which have no significant impact on the overall statistical results.

The major difference between Old Japanese and Modern Japanese word initials concerns /w/ which has disappeared before /i/, /o/ and /e/, and /y/ which has disappeared before /e/, as well as the relative increase of /b/, /d/, /z/ and /g/ as word initials. The other major change between Old Japanese and Modern Japanese is the evolution /p/ > /f/ > /h/, which will be discussed later in the paper.

did not contrast. For these reasons, the /h/ row of Table 2 represents the /h/ phoneme in modern Yamato Japanese as well as its ancestor /p/ in the archaic language.

Figure 6 recapitulates the order of segmental preferences at the beginning of C1 and C2 on the basis of the standard deviation observed for each initial in dvandva compounds.

Table 2. Frequency of initial segments in dvandva compounds constituents, in comparison to expected frequency.

	Corpus of Reference	Corpus of Dvandva Compounds							
		Constituent 1				Constituent 2			
	Observed	Exp.	Obs.		DR	Exp.	Obs.		DR
	%	*n*	*n*	%		*n*	*n*	%	
b	0.1	0.2	1	0.5	+0.8	0.2	3	1.5	+2.8
m	11.3	22.7	20	10.0	–2.7	22.7	21	10.4	–1.7
w	0.2	0.4	1	0.5	+0.6	0.4	2	1.0	+1.6
t	12.9	25.9	23	11.4	–2.9	25.9	17	8.5	–8.9
d	0.1	0.2	8	4.0	+7.8	0.2	1	0.5	+0.8
s	11.5	23.1	13	6.5	–10.1	23.1	21	10.4	–2.1
z	0.1	0.2	1	0.5	+0.8	0.2	0	0.0	–0.2
n	6.3	12.7	18	9.0	+5.3	12.7	5	2.5	–7.7
r	0.2	0.4	0	0.0	–0.4	0.4	0	0.0	–0.4
k	17.0	34.2	17	8.5	–17.2	34.2	58	28.9	+23.8
g	0	0	0	0.0	0	0	1	0.5	+1
y	6.3	12.7	21	10.4	+8.3	12.7	10	5.0	–2.7
h	11.2	22.5	12	6.0	–10.5	22.5	40	19.9	+17.5
’	22.8	45.8	66	32.8	+20.2	45.8	22	10.9	–23.8
Total	100	201	201	100	0	201	201	100	0
Standard Deviation					*8.83*				*10.70*

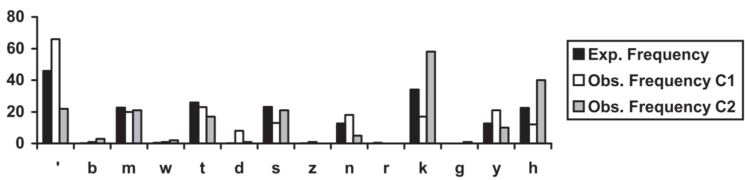

Figure 5. Segment frequency at the beginning of constituents in dvandva compounds of native and mixed origin (total=201 compounds).

a. C1 initials:

−													+
k	**h**	**s**	t	m	r	g	w	z	b	n	d	**y**	’

b. C2 initials:

−													+
’	t	n	y	s	m	r	z	g	d	w	b	**h**	**k**

Figure 6. Classification of dvandva compounds C1 and C2 initials, from least to most frequent, based on standard deviation (in bold: initials with a statistically significant occurrence rate).

3.1.2. Comments and discussion

Considering, as we did for CAL forms, that only those initials with a deviation higher in absolute value than the standard deviation for a given constituent are statistically significant, we notice the following anomalous frequencies.

For C1, the standard deviation is 8.83. /’/ occurs 66 times as an initial, that is, 32.8%, when we expect only 22.8% (dr for /’/=+20.2), and /y/ occurs 21 times, or 10.4%, when 6.3% are expected (dr=+8.3). By contrast, /k/, which appears only at the beginning of 17 C1’s (8.5%, when we would expect 17%), is statistically the least favored C1 initial (dr for /k/=−17.2). At the beginning of C1, two other segments, /s/ and /h/, occur significantly less than expected. There are 6% of /h/-initial C1’s when 11.2% are expected (dr=–10.5), and 6.5% of /s/-initial C1’s when 11.5% are expected (dr=–10.1).

For C2, the opposite tendency emerges. The standard deviation rate is 10.7, and /k/ is by far the preferred C2 initial, with an occurrence rate of 28.9%, representing 58 items (dr=+23.8), when 17% are expected. It is followed by /h/, 19.9% with 11.2% expected, representing 40 occurrences (dr=+17.5). On the contrary, /’/ is clearly disfavored in this position, with only 10.9% of /’/-beginning C2’s (22 items, dr=–23.8) when 22.8% are expected.

To sum up, /k/, /h/ and /s/ are statistically underrepresented at the beginning of C1, while /'/ and /y/ occur more than expected. As C2 initials, /k/ and /h/ are the two segments which display a significantly high frequency, while /'/ is statistically rare, and exhibits the greatest discrepancy between its C1 and C2 frequencies.

Let us consider now how the initials preferences in dvandva compounds constituents fit the SFH principles discussed in section 2.4. In dvandva compounds, /'/ appears as the favorite C1 initial, as it did in CALs' C1s, and this preference can be likewise attributed to its weakest-of-all-segments status in the SFH. Concerning /y/, note that even if /y/ did not rank as a statistically significant frequent CAL C1 initial, its preference as a C1 initial in dvandva compounds conforms to the SFH hypothesis since /y/, as a semi-consonant, is one of the weakest segments after /'/.

At the beginning of C2, observe that /k/ is the most frequent initial, as it was in CALs, in accordance with the SFH. The presence of /h/ as a significantly frequent C2 initial might seem unexpected at first sight because the SFH principles do not predict that a glottal fricative would rank second to /k/. Recall that at the beginning of CALs, /p/, rather than /h/, was the second most favored initial. But recall also that due to the fact that /p/ is not a licit word-initial in Yamato Japanese, no dvandva compound constituent starts with /p/. Taken together, these two facts (overrepresentation of /h/ as a C2 initial and absence of /p/ word initially) make sense. This is because, as already mentioned, Modern Japanese /h/ in Yamato and Sino-Japanese words was formerly */p/ and turned into /h/ word-initially before the 10th century. As most Yamato dvandva compounds used in Modern Japanese were presumably created at the stage of pre-Old Japanese, we can account for the fact that /h/ ranks second after /k/ as a C2 initial in dvandva compounds. At the time the compound was coined, the initial consonant of C2 was /p/, not /h/. Note that /h/ also exhibits a significant discrepancy in its C1 and C2 frequency, since it is between five and six times more frequent at the beginning of C2 than at the beginning of C1. Consequently, the greater frequency of /h/ as a C2 initial in modern

Yamato and mixed dvandva compounds simply reflects the older value of /h/ as */p/. Hence, the fact that /h/ occurs more than expected at the beginning of C2 does not contradict the SFH hypothesis. On the contrary, it constitutes further evidence for its validity, and proves that the segmental preferences uncovered in CALs and dvandva compounds are old and firmly established in the language. Concerning /s/, whose frequency at the beginning of C1 is significantly low, the results are somewhat unexpected since /s/, as a coronal fricative, is not supposed to rank as high in the SFH as /k/, /p/ or /t/. But if, as some scholars assume, Modern Japanese /s/ was formerly an affricate, either */ts/ (Arisaka, 1936) or */tʃ/ (Kobayashi, 1981), the puzzle finds a natural solution: since they start with a voiceless stop consonant, the affricates */ts/ and */tʃ/, which can also be considered as relatively marked phonologically because they are contour segments, stand high in the SFH. This explains why they were disfavored as a C1 initial at the stage of pre-Old Japanese in dvandva compounds.

3.1.3. Why *are-kore* ("that and this") rather than *kore-are* ("this and that"): a reply to Kageyama (1982)

Segmental preferences at the beginning of dvandva compounds' constituents strikingly resemble those of CALs, but they are also of a more general significance for Japanese linguistics. It is generally assumed that in dvandva compounds, the order of the constituents is determined by semantic or pragmatic factors, and that the member with highest priority tends to come first (Kageyama, 1982). Even if a number of examples in our sample confirm this assumption, it is noteworthy that resorting to semantic or pragmatic factors alone fails to account for the figures in Table 2. Our data show that besides pragmatic or semantic ones, phonological factors are also relevant in accounting for constituent order in Japanese dvandva compounds (see also Murata, 1993, for a similar claim, albeit based on a different sound hierarchy). The best illustration of this principle can be found in the deictic expressions *are-kore* "that (and) this", *achira-kochira* "in that direction (and) in this direction", and so on. In these words, and contrary to what could be expected on the basis of pragmatic factors, the term referring to an object or place closest to the speaker comes second. This, we claim, is just a consequence of the fact that *kore* and *kochira*, although referring to the speaker, start with a strong segment contrary to *are* and *achira*. Thus, the ordering of the constituents in *are-kore* and *achira-kochira* should rather be seen as another instance of the melodic preferences uncovered in CALs, where phonology has precedence over semantics and pragmatics, rather than as being governed by the semantic notion of distance and non-egocentricity (Kageyama, 1982), or “farness” precedence (Murata, 1993).

3.2. Echo-words

There exist, within the Japanese mimetic strata, a subclass of compounds of the echo-word type. A majority of them exhibit imperfect reduplication by means of a change in the consonant and vowel of the initial syllable of the second member such as *mecha-kucha* “incoherently”, *teki-paki* “briskly”, *maze-koze* “pell-mell” or *dota-bata* “noisily”. A few examples which do not involve reduplication also occur, for instance *gata-pishi* “rattlingly”, as well as five forms where only the vowel of the initial C2 syllable is modified (*kara-koro* “with a clatter”). Although echo-words are characterized by a number of specific properties which are due to their sound symbolic nature (see Hamano, 1998, Tsuji, 2003, for detailed studies), they nevertheless exhibit segment organization and preference tendencies which can be compared to those already uncovered in CALs and dvandva compounds.

3.2.1. Method and data

The methodology employed for CALs and dvandva compounds was applied to a corpus of 97 echo-words adapted from the list provided by Tsuji (2003). As a corpus of reference, we retained the 719 dissyllabic mimetic roots provided by the same author. These roots are all of the form (C)VCV, and they are instantiated in modern Japanese mimetic words under different morpho-phonological patterns such as *koso-koso*, *kosori*, *kossori*, etc. for the root **koso* “secretly”. Once more, we observe striking differences, resembling those already uncovered in CALs and in dvandva compounds with respect to segmental preferences at the beginning of C1 and C2. The results are shown in Table 3 and Figure 7.

Table 3. Frequency of initial segments in echo-word constituents, in comparison to expected frequency.

	Corpus of Reference		Corpus of Echo Words							
			Constituent 1				Constituent 2			
	Observed		Exp.	Obs.		DR	Exp.	Obs.		DR
	n	%	*n*	*n*	%		*n*	*n*	%	
p	54	7.5	7.3	3	3.1	–4.3	7.3	4	4.1	–3.3
b	50	7.0	6.7	4	4.1	–2.7	6.7	5	5.2	–1.7
m	61	8.5	8.2	14	14.4	+5.8	8.2	8	8.2	–0.2
w	15	2.1	2.0	3	3.1	+1	2.0	0	0.0	–2.0
t	59	8.2	8.0	15	15.5	+7	8.0	3	3.1	–5.0
d	46	6.4	6.2	9	9.3	+2.8	6.2	1	1.0	–5.2
s	67	9.3	9.0	6	6.2	–3	9.0	12	12.4	+3.0
z	40	5.6	5.5	4	4.1	–1.4	5.5	0	0.0	–5.4
n	40	5.6	5.5	9	9.3	+3.6	5.5	0	0.0	–5.4
r	2	0.3	0.3	0	0	–0.3	0.3	0	0.0	–0.3
k	72	10.0	9.7	3	3.1	–6.7	9.7	51	52.6	+41.3
g	82	11.4	11.1	7	7.2	–4.1	11.1	3	3.1	–8.1
y	20	2.8	2.7	3	3.1	+0.3	2.7	0	0.0	–2.7
h	60	8.3	8.1	5	5.2	–3.1	8.1	9	9.3	+0.9
’	51	7.1	6.9	12	12.4	+5.1	6.9	1	1.0	–5.9
Total	719	100	97	97	100	0	97	97	100	0
Standard Deviation						*3.99*				*11.40*

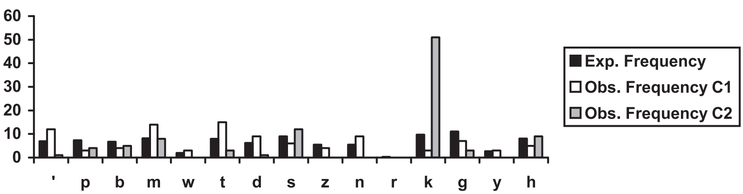

Figure 7. Segment frequency at the beginning of echo-word constituents (total=97 compounds).

Figures 8 shows the order of segmental preferences at the beginning of C1 and C2 on the basis of the standard variation observed for each echo-word constituent initial.

a. C1 initials:

–														+
k	**p**	**g**	h	s	b	z	r	y	w	d	n	**’**	**m**	**t**

b. C2 initials:

–														+
g	’	z	n	d	t	p	y	w	b	r	m	h	s	**k**

Figure 8. Classification of echo-words’ C1 and C2 initials, from less to more frequent, based on standard deviation (in bold: initials exhibiting a statistically significant frequency).

3.2.2. Comments and discussion

Considering C1 first, observe that there are six segments which exhibit statistically anomalous frequency rates at the beginning of C1. Standard deviation for C1 amounts to 3.99. The velar and labial stops /k/, /p/ and /g/ are clearly disfavored while /t/, /m/ and /’/ are preferred. /k/ is the least favored C1 initial, with a dr=–6.7, that is 3 occurrences (3.1%). It is followed by /p/, dr=–4.3, which is found at the beginning of 3 C1’s (3.1%), and by /g/, dr=–4.1, which appears 7 times as a C1 initial (7.2%). Segments which display a positive deviation rate at the beginning of C1 are /t/, which occurs 15 times as an initial (15.5%, dr=+7), /m/, which occurs 14 times (14.4%, dr=+5.8) and /’/, 12 times (12.4%, dr=+5.1).

It does not come as a surprise to find, once more, that /k/ is the least favored C1 initial, followed by /p/ and /g/, which are also situated at the top of the SFH. However, the ranking of /t/ as the most favored C1 initial is unexpected. Although a coronal, /t/ is a voiceless stop which should rank lower than /’/ as a C1 initial. Similarly, /m/, as a labial nasal stop, can be considered as a high ranking consonant in the SFH, and it is not expected to occur so frequently at the beginning of C1[11]. But note however that although /’/ is not the most favored C1 initial in echo words, it still does occur significantly more frequently than expected in that position, just as it does in CALs and in dvandva compounds.

The standard deviation for C2 is 11.4. At the beginning of C2, /k/ is the only consonant whose frequency is significantly higher than expected: 51 occurrences. More than half of all echo-words have a C2 starting with /k/ (52.6%, dr=+41.3)[12)]. This extreme scoring of /k/ causes the standard deviation to be relatively high with respect to the total number of items in the corpus and somewhat distorts the general appreciation of the figures, since outside /k/, the segment occurring most frequently as a C2 initial, the fricative /s/, occurs 12 times. As a consequence, no segment other than /k/ exhibits a deviant frequency rate as a C2 initial. However, notice that the two initials /g/ and /'/ stand at the bottom of the scale for C2, and although their deviation is not statistically anomalous (for the reasons just exposed), they still deserve some comments. This time, it is /g/ which is the least favored segment, just before /'/, according to our deviation rate based classification. We already saw that /g/ was also disfavored as a C1 initial. Actually, the low score of /g/ in our survey of echo-words' phonemic organization is a direct consequence of the special status of /g/ in mimetic roots since, as the corpus of reference figures show, /g/ is—quite surprisingly given its status in the Japanese phonological system—the most frequent segment at the beginning of Japanese mimetic roots (see also Hamano, 1998 for a similar observation). So it is actually this overrepresentation of /g/ at the beginning of mimetic roots, rather than its attendant relative underrepresentation at the beginning of echo-word constituents, which should require an explanation. Concerning /'/, there exists a major bias between the C1 frequency and the C2 frequency: /'/

11) Following some comments by two reviewers, we acknowledge that the choice of /t/ and /m/ as echo-words' C1 favorite initials does not fit the SFH hypothesis. This issue requires further investigation, that we leave to future research.

12) Quite surprisingly, the use of /k/ as a replacing segment in echo-word formation is not completely fossilized in Japanese: we recently recorded the example *mashen-kashen* (based on the French word *machin* "thing, stuff"), with the meaning "all that stuff whatever it is", which was spontaneously produced by a French-speaking native Japanese speaker.

appears 12 times at the beginning of C1, but only once at the beginning of C2, this unbalance being quite similar to what we already observed for CALs and dvandva compounds. We can therefore safely conclude that /'/ is clearly avoided as a C2 initial in echo-words.

/k/ is actually the sole consonant which exhibits a truly remarkable distributional patterning, with, on one hand, its deviation rate as a C2 initial far above that of the standard deviation, and, on the other hand, its enormous frequency disparity between C1 and C2: while /k/ appears as a C2 in 51 echo-words, it occurs only three times at the beginning of C1.

To conclude, the phonotactic tendencies observable in dvandva compounds and in echo-word compounds closely resemble those occurring in CALs, especially with respect to the scoring of /k/ and /'/ as C1 and C2 initials. This cannot be the result of mere chance. On the contrary, it shows that the SFH is a pervasive principle in Japanese morpho-phonemics. Moreover, the fact that all three /k/-starting echo-words' C1's also have /k/ as their C2 initial (*kata-koso, kara-koro, kasa-koso*) and that the only /'/-starting echo-word C2 has /'/ as its C1 initial (*echi-ochi*) also provides clear evidence that, at least in echo-words, the phonemic quality of C1 and that of C2 are linked through the SFH: C1 should not be stronger than C2, so if C1 is strong (that is, starts with /k/), its C2 must be at least as strong (that is, start with /k/), and if C2 is weak (starts with /'/), its C1 must not be stronger (i.e. it will optimally start with /'/).

4. CALs, dvandva compounds and echo-words as non-headed compounds

As stated above, CALs, dvandva compounds and echo-words exhibit similar patterns of initial preferences at the beginning of their constituents. They especially favor /'/ at the beginning of C1 and /k/ at the beginning of C2, while conversely rejecting them at the beginning of the other constituent. We assume that it is because they are non-headed compounds that CALs, dvandva compounds and echo-words behave similarly with regard to the tendency to promote a strong segment at the beginning of C2 and a weak one at the beginning of C1.

That dvandva compounds and echo-words are semantically non-headed is obvious and uncontroversial. *Yama-kawa* “mountains and rivers”, is not a kind of *kawa* “river” and neither the C1 *yama* nor the C2 *kawa* is the head of the compound (compare with the right-headed compound *yama-gawa* “mountain river”, which refers to a kind of river, and in which *rendaku* occurs). The same can be said of *mecha-kucha* “messy”, which is not a kind of *kucha.* Identical reasoning applies to CALs. A *paso-kon* is not a kind of *kon*, since *kon* does not exist as an autonomous word. Although the base *paasonaru konpyuutaa* is right-headed, both semantically and morphologically, its shortened form *paso-kon* is no longer interpreted as a compound on the semantic level. It has become a semantically simplex form through the clipping process.

Further evidence in support of the non-headedness of CALs, dvandva compounds and echo-words can be found in the fact that they share two additional properties which are characteristics of non-headed compounds in Japanese.

First, CALs, dvandva compounds and echo-words do not undergo *rendaku,* the morpho-phonological process by which the initial obstruent of the second member of a compound is voiced in order to materialize the fact that two words undergo composition, as in *yama+sakura* > *yama-z̲akura* “wild cherry tree”. *Rendaku* is a complex phenomena (see Satô, 1989, Itô & Mester, 2003, among others), involving huge idiosyncrasy and variation, but it is to be noted that mimetic words, dvandva compounds and compound clippings (of any etymological stratum) constitute, each in their own right, clear and indisputable cases of compounds which never exhibit *rendaku*. It is important to insist on the fact that it is not just because of their Western origin that CALs do not undergo *rendaku*. It is principally because they are clipped forms, and no clipped form in Japanese, whether of Yamato, Chinese or Western origin, exhibits *rendaku*. *Rendaku* can be viewed as a process materializing the morphological and syntactic dependency of the initial constituent of a compound structure on the final one. As Itô and Mester (2003: 85) observe, “the basic rule is that only modifier-head compounds permit *rendaku* voicing”. Accordingly, compounds which do not permit *rendaku* are not modifier-head compounds, and since head-modifier compounds (i.e. left-headed compounds) do not exist in Japanese, compounds which do not permit *rendaku* are necessarily non-headed.

The preference for strong segments such as /k/ (and, in CALs and dvandva compounds, for /k/ and /p/) at the beginning of C2 can be viewed as the negative version of *rendaku*. *Rendaku* is often compared to some sort of phonic “paste” or “starch” by phonologists (Komatsu, 1981: 104), and it can be considered as a type of consonant lenition. In CALs, dvandva compounds and echo-words, what we have with the preference for a strong consonant at the beginning of the second constituent looks more like a “phonic separator” which serves to highlight the boundary between the two components by promoting segments with greater force, i.e. greater perceptibility and distinctiveness.

Second, CALs, dvandva compounds and echo-words are morphologically compound nouns or noun-like forms which constitute a single phonological word but do not exhibit the typical accentual pattern characteristic of compounds in Japanese. Whereas Japanese (right-)headed compound nouns are normally accented on the initial mora of the second member or the final mora of the first member, CALs, dvandva compounds and echo-words are not. CALs and echo-words are unaccented or initially accented, while dvandva compounds keep the accent of the first member in its original position. This can be interpreted as further evidence that CALs, dvandva compounds and echo-words are not right-headed compounds, because otherwise they would preserve accent location on the head member of the structure, i.e. the rightmost member, or put an accent on the last mora of their C2. That *rendaku* and accentuation entertain a number of intricate links is further documented by Satô (1989).

So, although totally distinct and unrelated etymologically, semantically and lexically, CALs, dvandva compounds and echo-words share three common properties which are related to their non-headedness: they present a tendency to favor strong consonants at the beginning of their second member, they do not undergo *rendaku*, and they do not behave accentually like other compound nouns.

However, there exists a major difference between echo-words and dvandva compounds on one hand, and CALs on the other hand. In CALs, the order of the two constituents is relevant, and is determined by semantic and syntactic rules, whereas it is not in echo-words and in dvandva compounds[13]. Take for instance the two English lexemes *family* and *computer*, both adopted into Japanese as autonomous words. The compound *famirii konpyuutaa* {family computer} was coined in Japan and refers to a kind of computer, designed for familial use, whereas a *konpyuutaa famirii* {computer family} would be a kind of family, whose members are computer fanatics. Therefore, while it seems reasonable to assume that the order of the constituents in dvandva compounds and echo-words is determined by comparing the positions of the initial segments of each member in the SFH, the member starting with the strongest of the two segments being attributed the final position, this cannot be so in CALs. This is because unabbreviated compound loanwords are headed structures, whose constituent order cannot be modified because this would entail a major modification in the semantic content of the form. And since CALs have to be faithful to their base, they cannot modify the ordering of their constituents, even though they have become non-headed. In addition, it is crucial to note that the output forms of CALs are not obtained through segmental modification of the base. Generally, it is not possible to change the segmental build-up of the base, for instance to insert /k/ at the beginning of a /'/-starting C2 (example *gurafikku ikoraizaa* > **gura-<u>k</u>iko* instead of *gura-iko*{graphic equalizer}), nor to select a portion other than the initial part of the second member of the base (example *gurafikku*

13) Note however that there exist a couple of examples of coordinated compound loans of the dvandva type, such as *purasu mainasu* {plus minus}, which yields *pura-mai*.

ikoraizaa > **gura-rai*)[14]. So the overwhelming majority of CALs preserve the segmental material present at the beginning of the base, and the relationship between the base and the CAL only involves truncation of the final part of the base constituents.

There is an important theoretical implication to this observation: it means that the possibility to derive a CAL is determined at the base level. This issue requires additional research (in particular, further data and analysis concerning the principles which govern loanword composition in Japanese are needed[15]), but one can imagine that if a compound base does not have the right phonotactic profile, for instance if its second member starts with /'/, it will be preferable to avoid CAL derivation[16] rather than to produce an ill-formed CAL. On the contrary, if a base already fulfills the preferred phonotactic requirements, that is, ideally, if its second member starts with /k/ or /p/ and its first member with /'/ or /n/, it will be more likely to undergo truncation. Thus not all bases have the same aptitude to produce a CAL, and the likelihood for a given base to produce a CAL depends on its phonotactic organization. One could invoke here some sort of Darwinian segmental determinism and natural selection.

However, note that there is no absolute ban on segmentally non-optimal CALs. Although rare and statistically underrepresented, forms such

14) As stated in footnote 7, a small number of forms which copy a non initial part of the second member of the base, as well as the CAL *ban-kame*, in which the initial constituent's final consonant /k/ has been syllabified as the C2 initial, can be found. However, these are marginal cases with no significant impact on the overall statistics, so that non initial anchoring of the C2 cannot be considered a regular strategy to avoid /'/-beginning C2s.

15) In particular, as one reviewer has pointed out, one should also check whether phonemic preferences similar to those discussed in this paper are enforced at the beginning of non-abbreviated compound loanwords constituents, that is, for instance, whether /k/-initial words such as *kaado*, *kuriimu* or *keesu* are more often used as a second member than as an initial member in such formations.

16) In such cases, one can suppose that other morphological devices are used to derive a shortened form. Two other productive processes are available in Japanese: elliptic derivation, which consists in copying integrally one of the lexemes of the base and deleting all the others (example *rosu* from *rosu anjeresu* {Los Angeles}, and alphabet-based acronymic derivation (*OL* from *ofisu redii* {office lady}).

as *furu-oke* (< *furu ookesutora* {full orchestra}) or *gura-iko* (< *gurafikku ikoraizaa* {graphic equalizer}) do occur. Thus CAL productivity is gradient, and it is correlated to the phonemic structure of the base. The same can be stated about dvandva compounds, since forms like *haru-aki* exist. Things are different with echo-words, because there is no instance of a form in which the C1 initial is stronger than the C2's.

5. Conclusion and further issues

In this paper we have documented the major phonemic properties of three different types of non-headed binary compounds in Japanese: CALs, dvandva compounds and echo-words. On the descriptive level, we have shown that phonemic principles crucially operate in determining the well-formedness of a compound word. The statistically significant segmental preferences and rejections observed at the beginning of C1 and C2 in CALs, dvandva compounds and echo-words are summarized in Table 4.

Table 4.

	C1		C2	
	+frequent	-frequent	+frequent	-frequent
CALs	n, **’**	**k**	**k**, p	**’**
Dvandva Compounds	**’**, y	**k**, h (=*p), s	**k**, h (=*p)	**’**
Echo-words	t, m, **’**	**k**, p, g	**k**	

We saw that in all three classes, at the beginning of C1, the zero initial /’/ (i.e. the vocalic initial) is consistently preferred, while /k/ is consistently rejected.

At the beginning of C2, /k/ is always the most favoured segment, while /’/ is rejected in CALs and dvandva compounds (in the case of echo-words, as already stated, the figures for /’/ do not appear as statistically significant due to the bias caused by /k/, but remember that an indisputably important discrepancy does exist between the occurrence rate of /’/ as a C1 and C2 initial since /’/ was 12 times less frequent as a C2 than as a C1 initial).

Moreover, /p/, as well as /h/, its modern reflex in Yamato words, behave similarly to /k/ insofar as they are clearly disfavoured as a C1 initial in dvandva compounds and echo-words, and favoured as a C2 initial in CALs and in dvandva compounds. Segments /n/, /y/, /t/, /m/, /s/ and /g/ also sporadically exhibit anomalous frequencies but each of these segments appears only once in the table, so things are not as general and pervasive as what occurs with /'/, /k/ and /p/~/h/.

On the explanatory level, we have argued that phonemic preferences occurring in CALs, dvandva compounds and echo-words fall out from a general Segment Force Hierarchy, in which non coronal voiceless stops such as /k/ and /p/ are the strongest, and /'/, the weakest. The presence of a strong consonant at the beginning of a C2 and that of a weak one at the beginning of a C1 fulfills a demarcative and culminative function, which facilitates the recognizability of the lexical, morphological and phonological structure of the derived form.

In the particular case of CALs, the starting point of this study, we have seen that contrary to what is generally assumed, prosody alone is not enough to account for the derivation patterns. Melodic principles also operate, so that bases with non-optimal constituent initials are less likely to undergo truncation than bases which fulfill the segmental preferences expressed in the SFH. Clearly, the phonemic structure of the base conditions the possibility to derive a CAL. Accordingly, possible CALs are not all equally probable, and it seems reasonable to suppose that some bases do not undergo truncation because of their melodic organization.

The data brought forth by the present study therefore provides substantive evidence that segmental phonology plays a greater role than what is generally assumed in Japanese word composition.

No attempt has been made to provide a formal account of the phenomenon, but the explanation we have proposed is compatible with a number of different theories and frameworks.

A number of important issues remain to be addressed. We briefly mention a few of these. First, more research is needed on the notion of "segmental force", and on how such segmental force should be calculated.

Second, it is to be noted that the requirement to have a strong element in the second member of a compound under certain conditions is rather well attested across languages. Consider for instance reduplications with fixed segmentism affecting the first consonant of the second member of a compound. They frequently involve strong segments such as labial or velar consonants: Tamil (*paampu-kiimpu* from {paampu} "snake" Keane, 2001), Khalkha Mongolian (*japon-mapon* from {Japon} "Japan", Kubo, 1997), Basque (*handi mandi* from {handi} "big", Lafitte, 1978), Turkish (*çocuk mocuk* from {çocuk} "child"), Russian (*sifilis-pifilis* from {sifilis} "syphilis" Waugh & Jakobson, 1979), etc., or marked segments combinations such as *shm* in English (*table-shmable*). Moreover, Cooper & Ross (1975) argue that second position constituents in English irreversible compounds and word phrases contain a more obstruent initial segment, a claim which has been confirmed by subsequent studies (Pinker & Birdsong, 1979, Wright, Hay & Bent, 2005). According to these authors, such expressions tend to place the longest or phonetically "heaviest" item at the end (see also Hagège (1985), who speaks of the "loi du second lourd", and Malkiel, 1959, for a review of the formal factors operating in the ordering in binomial compounds).

Third, it would be necessary to investigate more thoroughly the interaction between phonological constraints and semantic or pragmatic ones. In dvandva compounds, semantic considerations sometimes override phonological ones (examples *haru-aki* "spring and autumn", or *yoshi-ashi* "good and evil"), i.e. the positive connotation or higher pragmatic priority can cause an element with a strong initial to come first in the compound (Kageyama, 1982, Murata, 1993, Malkiel, 1959). Similarly, one can suppose that in CAL derivation, a form which is not optimal at first sight might be created if it is really needed for pragmatic reasons, in spite of its poor segmental make-up. This is the reason why CALs such as *kaa-nabi* (< *kaa nabigeetaa* {car navigator}) or *kame-riha* (< *kamera rihaasaru* {camera rehearsal}) do sporadically occur.

Fourth, there is evidence that phonemic preferences might involve segments occurring in positions other than the initial one in a given constituent. For instance, C2's starting with a labial or /h/, and whose second consonant is a coronal other than /r/, are extremely rare in our corpus. Such forms are generally avoided by deriving a monomoraic C2 instead of a bimoraic one, as the examples in (3) show. They have to be compared with the examples in (4), which possess a C2 starting with a labial followed by /r/ as second consonant, and with (5), whose C2 does not start with a labial or /h/.

(3)

rabu hoteru	> *rabu-ho*	**rabu-hote*	{love hotel}
burakku bisuketto	> *bura-bi*	**bura-bisu*	{Black Biscuit}
purasutikku moderu	> *pura-mo*	**pura-mode*	{plastic model}
roiyaru hosuto	> *roi-ho*	**roi-hosu*	{Royal Host}
akosutikku gitaa	> *ako-gi*	**ako-gita*	{acoustic guitar} etc.

(4)

animeeshon parodii	> *ani-paro*	{animation parody}
ajiteeshon puropaganda	> *aji-puro*	{agitation propaganda}
don perinyon	> *don-peri*	{Dom Pérignon}
ekisaito burogu	> *eki-buro*	{Excite blog}
erotikku parodii	> *ero-paro*	{erotic parody} etc.

(5)

dorimu kyasutaa	> *dori-kyasu*	{dream caster}
hangaa sutoraiki	> *han-suto*	{hunger strike}
anarogu dejitaru	> *ana-deji*	{analog digital}
gasorin sutando	> *gaso-suta*	{gasoline stand}
depaato resutoran	> *depa-resu*	{department restaurant} etc.

One should also see whether vowel quality plays a role in the ordering of CALs and dvandva compounds, as it does in echo-words (Murata, 1984).

Moreover, building on the evidence uncovered in the present paper, it would be interesting to check whether similar (or opposite) segmental preferences occur in other types of Japanese compounds such as Yamato right-headed compounds or Sino-Japanese compounds.

Lastly, another point of interest is whether phonemic principles are also relevant in accounting for other types of loanword truncations, namely back truncations and front truncations such as *guriin hausu* > *hausu* {greenhouse}, *paato taimu* > *paato* {part time}, *rosu anjeresu* > *rosu* {Los Angeles}, which, according to Nishihara et al. (2001), account for 53% of all truncations in Japanese loanwords. There are good reasons to assume that the phonological nature of each constituent initial plays a role, but a more systematic study is needed to confirm this insight.

References

Arisaka Hideyo (1936) Jôdai ni okeru sagyô no on'in. Reprinted in: *Kokugo on'inshi no kenkyû* (1969), 145–159. Tôkyô: Sanseidô.

Bat-El, Outi (1996) Selecting the best of the worst: the grammar of Hebrew blends. *Phonology* 13-3: 283–328.

Cooper, William E. and John. R. Ross (1975) World order. In: *Papers from the parasession on functionalism*, 63–111. Chicago: Chicago Linguistic Society.

Cutler, Anne and William E. Cooper (1978) Phoneme-monitoring in the context of different phonetic sequences. *Journal of phonetics* 26: 221–225.

Fougeron, Cécile (1999) Prosodically conditioned articulatory variations: a review. *UCLA Working papers in phonetics* 97: 1–74.

Hagège, Claude (1985) *L'homme de paroles*. Paris: Fayard.

Hamada, Atsushi (1954) Hagyôon no mae no sokuon—p-on no hassei. *Kokugogaku* 16. Reprinted in: *Zoku chôsen shiryô ni yoru nihongo kenkyû* (1983), 71–80. Kyôto: Rinsen Shoten.

Hamano, Shoko (1998) *The Sound symbolic system of Japanese*. Stanford & Tôkyô: CSLI Publications & Kuroshio.

Hashimoto, Shinkichi (1928) Hagyô shiin no hensen ni tsuite. Reprinted in: *Kokugo on'in no kenkyû* (1950), 29–45. Tôkyô: Iwanami Shoten.

Hayashi, Chikafumi (1992) Hagyô ten'on ha naze Heian jidai ni okotta ka. *Kokugo to kokubungaku* 69-11: 110–119.

Hooper, Joan, B. (1976) *An Introduction to natural generative phonology*. New York: Academic Press.

Itô, Junko (1990) Prosodic minimality in Japanese. In: Deaton K. et al. (eds.) *Parasession on the syllable in phonetics and phonology*, CLS 26, volume 2: 213–239. Chicago: Chicago Linguistic Society.

———— and Armin Mester (2003) *Japanese morphophonemics – markedness and word structure*. Cambridge, MA: MIT Press.

Jidai Betsu Kokugo Daijiten (1967) Tôkyô: Sanseidô.

Kageyama, Tarô (1982) Word formation in Japanese. *Lingua* 57: 215–258.

Keane, Elinor (2001) *Echo words in Tamil*. Phd dissertation, Oxford: Merton College.

Kobayashi, Akemi (1981) Ennin no kijutsu suru sansukuritto onsetsu *ca* no onka. *Ôsaka Gaikokugo Daigaku gakuhô* 52: 63–80.

Komatsu, Hideo (1981) *Nihongo no on'in. Nihongo no sekai 7*. Tôkyô: Chûô Kôronsha.

Kubo, Tomoyuki (1997) Reduplication meduplication in Khalkha Mongolian. *Gengo Kenkyû* 112: 66–96.

Labrune, Laurence (2002) The prosodic structure of simple abbreviated loanwords in Japanese: a constraint-based account. *Onsei kenkyû* 6-1: 98–120.

Lafitte, Pierre (1978) *Grammaire basque*. Donostia/San Sebastian: Elkar.

Malkiel, Yakov (1959) Studies in irreversible binomials. *Lingua* 8: 113–160.

Motvani, Prem (1991) *Nichijô gairaigo yôhô jiten/a Dictionary of loanwords usage katakana—English*. Tokyo: Maruzen.

Murata, Tadao (1984) Jinkô onomatope ni yoru nihongo onsei haierâki. *Gengo Kenkyû* 85: 68–90.

———— (1993) Nichi-eigo no ABgata onomatope—jûfukukei, tôi kôzô hyôgen no kankei. In: *Onomatopia gion gitaigo no rakuen*, 101–125. Tôkyô: Keisô Shobô.

Nishihara, Tetsuo, Jeroen van de Weijer and Kensuke Nanjo (2001) Against headedness in compound truncation: English compounds in Japanese. In: J. van de Weijer and T. Nishihara (eds.) *Issues in Japanese phonology and morphology*, 299–324. Berlin/New York: Mouton de Gruyter.

Paradis, Carole and Jean-François Prunet (eds.) (1991) *The Special status of coronals, internal and external evidence.* Phonetics and phonology 2. San Diego: Academic Press.

Pinker, Steven and David Birdsong (1979) Speakers' sensitivity to rules of frozen word order. *Journal of Verbal Learning and Verbal Behavior* 18: 497–508.

Prince, Alan, and Paul Smolensky (1993) *Optimality theory, constraint interaction in generative grammar*. Ms, Rutgers University, New Brunswick.

Satô, Hirokazu (1989) Fukugôgo ni okeru akusento kisoku to rendaku kisoku. In: *Kôza nihongo to nihongo kyôiku* 2, 233–265. Tôkyô: Meiji Shoin.

Selkirk, Elisabeth (1984) On the Major Class Features and Syllable Theory. In: Mark Aronoff and Richard Oehrle (eds.) *Language sound structure*, 107–136. Cambridge: MIT Press.

Tsuji, Sanae (2003) *Les impressifs japonais*. Lyon: Presses Universitaires de Lyon.

Waugh, Linda and Roman Jakobson (1979) *The Sound shape of language.* Brighton: Harvester Press.

Wright, Saundra K., Jennifer Hay and Tessa Bent (2005) Ladies first? Phonology, frequency, and the naming conspiracy. *Linguistics* 43-3: 531–561.

日本語ノンヘデッド複合語における音素上の偏り
「ワープロ」,「あれこれ」,「めちゃくちゃ」の共通点について

ローランス・ラブリュヌ

（ボルドー第3大学）

本稿は日本語における外来複合語の短縮形（例：ポケモン），エコーワード（例：めちゃくちゃ）および対立語（例：白黒）といった3種類の2項からなるnon-headed複合語を対象に，これらの語の音韻構造，特に前項および後項における頭の音素の分布と頻度を統計的に検討する．その結果として, 3種類の複合語の適格性は，各項の頭の音素の性質によって決定されるということを明らかにする．統計的に有意な最も際立った特徴として，ア行（つまりゼロ子音）で始まる要素は複合語の第2項として現れることが非常に少なく，カ行で始まる要素は第2項として最も現れやすいということが言える．本稿においては，これらの音素上の偏りは，知覚的および音韻的な要因に規定された音力階層によって説明され，またこのような複合語の音韻構成は機能的な動機付けによっていると主張する．

（受理日　2005年10月4日　　最終原稿受理日　2006年2月3日）