Consonant variation within words

Joost van de Weijer
Department of Linguistics and Phonetics
Lund University
Helgonabacken 12
22362 Lund, Sweden
email: vdweijer@ling.lu.se

1. Introduction

The aim of this paper is to show that the consonants that a simple (monomorphemic) word is constructed with usually are different from each other. To illustrate this, consider words like *glass*, *drink*, *butter*, or *tray* in which no consonant occurs more than once (it is the pronunciation I am concerned with, not the spelling). Changing one consonant so that it becomes identical to another one would result in rather odd-sounding new words, e.g., *glagg*, *krink*, *butteb* or *bubber*.

Of course I realize that there are a great deal of words in which the same consonant occurs more than once, as there are *clock*, *deed*, *paper*, *text*. However, I believe that these words are in a minority, and of relatively low frequency so that they do not show up often in everyday speech.

I do not think that the phenomenon is restricted to any language in particular, but instead applies to many languages, although I would not dare to claim all. I am not aware of any study in which this specific claim was questioned, but there are a number of findings that point in the direction of the idea that identical consonants repulse rather than attract one another.

First of all, it is common that suffixes are deleted when the adjacent consonant in the stem is homophonous, a phenomenon that has been labeled morphological haplology (Stemberger 1981). In English, for example, the genitive *–s* is not realized in plural forms, unless the plural does not end on *–s* (cf. *the *boys's books* with *the children's books*). Similarly, in Swedish, the present tense markers *–er* or *–ar* are not realized when the verb stem ends on *–r*, e.g., *jag kommer*, (I come), but *jag hör* — instead of **jag hörer* (I hear).

Second, identical nearby consonants sometimes undergo change during a process of borrowing between languages or as a result of historical sound change, a process known as phonological dissimilation (Hock and Joseph 1996). Examples of such a change in loan words are the French word *marbre* which became *marble* when adopted into English, or the Italian word *tartuffeli* (truffle) which became *Kartoffel* (potato) in German. An example of an historical sound change is the Swedish word *nyckel* (key) which once was *lyckel*.

A third observation is that whereas relatively many languages show vowel harmony (a phonological process according to which vowels within a word become more similar), there are only few examples of languages with consonant harmony. A major exception being child language in which patterns of consonant harmony frequently occur (Smith 1973; Vihman 1978).

Finally, a well-known phonological constraint, the obligatory contour principle (OCP), states that 'adjacent identical segments are prohibited' (Clements and Hume 1995). Originally this constraint was applied to explain the distribution of lexical tones, but at a later stage it was also used to explain regularities in speech segments. McCarthy (1986), for instance, used the OCP to explain patterns in Arabic trilateral and quadriliteral roots that lack identical consonants.

In sum, it appears that there is a tendency for nearby consonants to be different. The aim of the present study is to take this claim one step further by stating that most morphologically simplex words are constructed of only different consonants. This means that not only nearby consonants are different, but also consonants further apart, as long as they are within the same morpheme. If this is true, then the distribution of consonants can be a valuable source of information for automatic morphological decomposition, since there is an increased likeliness that there is a morphological boundary between two identical consonants. A second, related, application is that the distribution of consonants may help the listener or a speech recognizer in identifying word boundaries. Since there are no reliable markers of word boundaries in spoken language as there are in written language (Cole and Jakimik 1980), many researchers have searched for cues that may aid in the recognition of spoken words (e.g., Cutler and Norris 1988; McQueen 1998). The presence of identical consonants, then, may be a simple cue for the listener to hypothesize that there is a word boundary between them.

As a starting point for investigating these research claims, I analyzed the distribution of identical consonants in a selection of Swedish words. Swedish is a language for which I believe that the hypothesis is true. It has a phonological structure that is in a number of ways similar to that of English. Its syllable structure, for example, is comparable to that of English, with consonant clusters consisting of up to three consonants in syllable-initial position and clusters up to three consonants in syllable-final position, or even five if suffixes are included (Sigurd 1965). The phonemic inventory contains 18 different consonants (Elert 1989), but additional consonants occur in loanwords, mainly from English or French.

I restrict my analysis to morphologically simple words, so that compounds, derivations and inflections are not included. I do not know of any phonological process (other than morphological haplology which only applies when the adjacent consonant is homophonous) according to which consonants in affixes change because the same consonant also occurs in the stem that the affix is attached to. The suffix *–s* for third person singular in English, for instance, is the same for verb stems containing an *s* (as in *he sits*) or any other consonant (as in *he hits*).

The remainder of the paper is organized as follows. In the next section I describe the characteristics of the material that was used, and how it was analyzed. Thereafter, I will address the following issues: How many words contain two or more identical consonants? Are certain consonants more likely to occur more than once within a word than others? Are words with identical consonants relatively common or rather rare?

2. Method

Material

The material was selected from a word list that was used for the production of a current Swedish pronunciation dictionary (Hedelin 1997). The complete list is composed of 116,362 entries. An entry consisted of the orthographic word form, a word-class code (noun, adjective, adverb, pronoun, verb, count noun, proper noun, preposition, article, interjection, abbreviation, conjunction, infinitive marker, prefix, word group), the word's pronunciation, and whether the word was a compound, a derivation or an inflected form.

Selection of the material

The aim of the selection was to obtain a word list of different monomorphemic words. Clearly, the dictionary contained a large number of entries that did not meet this requirement. Furthermore, a proportion of the dictionary was redundant for the purpose of the present analysis, since many entries were repetitions of the same word root (e.g., words that could be a noun as well as a verb). Luckily, a large proportion could be excluded on the basis of the information already provided in each entry. A total of 90,463 entries coded as compounds, e.g., *bordduk* (table cloth), derivations, e.g., *alkoholist* (alcoholic – noun), *alkoholism* (alcoholism), *alkoholisk* (alcoholic – adjective), etc., and inflected forms (e.g., present tense verbs, genitive forms of pronouns, etc.) were excluded automatically.

Obviously, it is debatable where the line between what counts as a morphologically complex word and a morphologically simple word should be drawn (see, e.g., Sproat 1992 for further discussion on this issue). From the list that resulted after the exclusion, I manually excluded an additional 1,164 of entries that I considered morphologically complex as well. The entries that were excluded contained words with a meaning that could easily be derived from the individual meanings of the components. These words included for instance a number of change of state verbs which in Swedish are constructed of an adjective plus the suffix *–na* (e.g., the verb *mjukna* (to become soft) consists of the adjective *mjuk* (soft) followed by the suffix *na*). When the meaning could not be derived from the word's individual components – for instance the meaning of the word *protest* (protest) – the word was not excluded.

Although this strategy worked reasonably well for bisyllabic words, it became more difficult for words of three syllables or longer. For that reason, I decided to restrict the analysis to words of one and two syllables only, and excluded all words that were longer than two syllables.

Finally, proper names (e.g., *Alfa Romeo, Charlotte*, *Amsterdam*), abbreviations and acronyms, (e.g., *SAS*, *NATO*), interjections, (e.g., *oj* (oops), *usch* (ugh)), orthographic variants (e.g., *mej* (me) instead of *mig*, or *sebra* (zebra) instead of *zebra*), and repeated entries of the same word root (e.g., *bank* in the meaning of 'sofa' as an alternative to *bank* in the meaning of 'bank') were also excluded. The final selection then consisted of 8,887 words. The number of consonants in each word varied from zero to eight.

Word Frequencies
Word frequency information was obtained from a language corpus collected at the University of Gothenburg. The research group there provided me with a list of approximately 100,000 word types (100,998) that had a frequency of 20 or more in a large corpus consisting mainly of newspaper articles. The sum of the frequencies of these words was somewhat below 57 million tokens (56,916,383). In this list, there were a total of 5,388 word types that also occurred in the selection of morphologically simple words. Together these had a total token frequency of little below 29 million (28,845,117).

Analysis
The decision whether a word contained two or more identical consonants was made on the basis of the transcripts provided in the dictionary, with the following two exceptions.

In the dictionary a distinction is made between long and short consonants. This difference is primarily context dependent, i.e., long consonants follow short vowels, as in the word *tack* /tAk˘/ (thanks), and short consonants follow long vowels, as in the word *tak* /tA˘k/ (roof). In addition, the difference in pronunciation is not realized in many Swedish dialects (Elert 1989). For these reasons, long and short variants of the same consonant were considered to be identical.

Second, in some dialects of Swedish, the /r/ has an effect on the place of articulation of one or more subsequent alveolar consonants. In these cases, the consonants adopt a retroflex or supradental place of articulation, and the /r/ disappears. The word *barn* (child), for example, is pronounced as /bA ˜/ (Elert 1989). The assimilated forms are given in the dictionary, but these were replaced in the analysis by the unassimilated form (e.g., /bA rn/).

Notice that these two criteria make the analysis more conservative, since the likeliness of finding two or more identical consonants within a word increases in both cases.

3. Results

During the analysis, it became clear that no consonant occurred twice within a syllabic onset or syllabic coda. For instance, there were no syllables that ended on *–tst* or *–ksk* or any other combination of three consonants. Although this is an intuitively plausible finding it is not a trivial observation since the presence of two identical consonants within a syllable onset or coda is in principle not prohibited by phonotactic constraints. In English, for instance, any noun ending on *–st* combined with a genitive *–s*, or any verb ending on *–sp* combined with third person singular *–s* results in words with codas with two identical consonants, e.g., *guest's* or *gasps*, etc.

A second thing that became clear during the analysis, was that no two consonants directly adjacent to each other ever were the same. This could have been possible, for instance, at the boundary between the first and the second syllable in a bisyllabic word. However, there were no such examples.

These two findings were the starting point for dividing the words according to their syllable structure as shown in Table 1. The left column of the table shows the structure of the words, starting with words that consisted of a vowel only. The O's and the C's in the following rows stand for Onset and Coda, where O1 and C1 refer to the onset and the coda of the first syllable, and O2 and C2 refer to the onset and the coda of the second syllable. C1 in the second row then, refers to all the words that had one or more consonants in the coda of the first syllable, but no consonants in the onset, or in the second syllable if there was any. The second column shows the numbers of word types with the structure listed in the first column. The third and the fourth columns show the absolute and the relative numbers of hits, i.e., the words that contained at least two identical consonants. The last column shows two examples of words with the structure given in column 1. The examples are, wherever possible, words with at least two identical consonants.

The table starts with five categories that could not have identical consonants. For the sake of simplicity, bisyllabic words with only one consonant in C1 and one consonant in O2 (for example the words *inte* and *syssla* in the table) are classified as if these two consonants both belonged to the onset of the second syllable. Together the first five categories consisted of 438 types, corresponding to 4.93% of the total.

The total number of words with at least two identical consonants was 1,001 (11.26%). Most of them contained no more than two consonants that were the same, a few common examples of which were the pronouns *någon* (someone) and *annan*, (other), the adverb *nästan* (almost), the conjunction *trots* (although), and the adjective/noun *svensk* (Swede, Swedish). However, there were 54 word types in which two consonants occurred twice, including the words *status* (status), *porträtt* (portrait), *struktur* (structure), and *taktik* (tactics). In addition, there were ten words in which one consonant occurred three times, including *skepsis* (skepticism), *substans* (substance), *traktat* (treaty), *census* (census). There were no words that contained more than three identical consonants or more than two consonants that occurred twice.

Table 1: Type count. *C* stands for coda; *O* stands for onset.

structure	*types*	*hits*	*(%)*	*examples*
–	3	0	(0.000)	*i* (in), *ö* (island)
C1	113	0	(0.000)	*upp* (up), *älv* (river)
C2	3	0	(0.000)	*oas* (oasis), *eon* (aeon)
O1	153	0	(0.000)	*trä* (wood), *vrå* (corner)
O2	166	0	(0.000)	*idé* (idea), *inte* (not)
C1O2	29	0	(0.000)	*önska* (wish), *ändra* (change)
O1O2	2,373	154	(6.490)	*syssla* (work), *kruka* (pot)
O1C1	2,432	185	(7.607)	*klipsk* (shrewd), *klok* (wise)
O2C2	424	45	(10.613)	*annan* (other), *essens* (essence)
O1C1O2	159	20	(12.579)	*virvla* (whirl), *dundra* (thunder)
O1C2	66	12	(18.182)	*nyans* (nuance), *neon* (neon)
O1O2C2	2,799	527	(18.828)	*banan* (banana), *staket* (fence)
C1O2C2	39	11	(28.205)	*emblem* (badge), *anstalt* (institution)
O1C1O2C2	128	47	(36.719)	*substans* (substance), *distrikt* (district)
totals	*8,887*	*1,001*	*(11.264)*	

Among the words that contained identical consonants, there were four relatively small groups of words that suggested that identical consonants in some cases serve a specific purpose. These four groups contained words that are typically used by children, onomatopoetic verbs and nouns, a set of nouns and adjectives that all had a pejorative or distasteful connotation, and finally a set of bisyllabic mimetic ('flip-flop') words that consist of two syllables that are exact or nearly exact replications of each other (see Table 2 for examples).

Table 2: Specific functions of identical consonants.

child words	*onomatopoetic*	*phonestemes*	*'flip-flop' words*
mamma (mother)	*pipa* (chirp, squeak)	*knark* (drugs)	*tiptop* (tip-top)
pappa (dad)	*bubbla* (bubble)	*skurk* (villain)	*vigvam* (wigwam)
baby (baby)	*mummel* (murmur)	*strunt* (rubbish, trash)	*pingpong* (ping-pong)
pippi (dickybird)	*tuta* (hoot)	*snusk* (uncleanness)	*picknick* (picnic)
jojo (yoyo)	*nynna* (hum)	*skolk* (truancy)	*virrvarr* (crisscross)
bebis (baby)	*babbla* (blather)	*skunk* (skunk)	*sicksack* (zigzag)
vovve (dog)		*stursk* (insolent, impudent)	*mischmasch* (mishmash)
dada (nanny)		*smisk* (smack)	*gonggong* (gong)
		slisk (cloyingly sweet)	*snicksnack* (chatter)
		slusk (shabby person)	
		smask (smacking noise)	
		slask (wet garbage)	
		stursk (stubborn)	

Which consonants were most likely to occur more than once in a word? Table 3 shows how often each consonant occurred overall and how often it occurred as identical consonant. Comparing the two frequencies it shows that /t/, /k/ and /s/ were relatively likely to occur more than once, since their overall relative frequencies were much lower than their frequency as identical consonants. On the contrary, /l/ was not likely to likely to occur more than once since there was an almost 7% difference in how often it occurred overall and how often it occurred as identical consonant. The difference between the two frequencies was not more than 3% for all the other consonants.

Table 3: Consonant frequencies (%).

consonant	total relative frequency	frequency as identical consonant
d	5.022	2.395
f	2.998	0.958
g	3.417	2.011
h	1.499	0.000
j	2.929	0.383
k	9.068	14.464
l	10.644	3.831
m	5.276	4.789
n	7.787	7.184
p	4.832	4.693
r	12.831	13.027
s	11.928	20.881
t	10.444	19.253
v	3.413	2.490
Ó	1.008	0.000
˓	0.568	0.192
N	2.180	0.479
S	0.389	0.192
b	3.766	2.778

Table 4 shows the results of the token count. Comparing the values listed in Tables 1 and 4 reveals that the total percentage of words with identical consonants has decreased from 11.26 to 1.57. A major cause of this decrease is that the numbers of tokens in the first five rows are much higher in Table 4 than in Table 1. Included in these categories are a small number of very highly frequent function words, such as *och* (and), *att* (to), *en* (a). However, all the percentages in the other rows of Table 4 are lower than the corresponding percentages in Table 1 as well, indicating that the words with identical consonants were relatively rare.

Table 4: Token count. *O* stands for Onset, *C* stands for Coda.

structure	*types*	*tokens*	*targets*	*(%)*
–	3	1,738,874	0	(0.000)
C1	92	6,866,180	0	(0.000)
C2	1	241	0	(0.000)
O1	134	3,171,934	0	(0.000)
O2	123	928,972	0	(0.000)
C1O2	21	127,133	0	(0.000)
O1C1	1,796	11,079,424	164,988	(1.489)
O1O2	1,397	2,244,855	37,974	(1.692)
O1C2	28	15,063	295	(1.958)
O2C2	243	1,067,550	44,522	(4.170)
C1O2C2	20	47,768	2,079	(4.352)
O1C1O2	71	54,653	4,013	(7.343)
O1O2C2	1,402	1,475,306	194,242	(13.166)
O1C1O2C2	57	27,164	5,021	(18.484)
totals	*5,388*	*28,845,117*	*453,134*	*(1.571)*

4. Conclusions

The hypothesis of the present investigation was that monomorphemic words that contain identical consonants are rare. The results showed that 11.26 percent of Swedish monosyllabic and bisyllabic words contained at least two identical consonants. However, these words seem to be relatively uncommon and, consequently, will not be heard or read often. The estimate that I found based on a selection of more than 28 million word tokens, I estimated that only 1.57 percent contained identical consonants.

Taken together, the results show that identical consonants within words indeed constitute a dispreferred pattern in Swedish. Based on the result of the token count, it would be expect to find a word with two or more identical consonants only one in every 65 words. The first question that is open for future research is whether this is also the case in other languages. I expect to find similar percentages in languages with a phonological structure that is comparable to that of Swedish, for instance the other Germanic languages. More interesting comparisons will be with languages that have deviant phonological structures, e.g., different syllable structure or fewer consonants.

A second question that needs to be answered is whether listeners are aware that identical consonants within words are relatively rare, and whether they use the presence of identical consonants for morphological decomposition. If this is true, then it is predictable that listeners will identify non-words with two identical consonants more quickly than non-words with only different consonants. A second prediction is that listeners will decompose morphologically complex words that contain identical consonants (e.g., *sits*) more quickly than words that do not (e.g., *hits*). I intend to test these predictions in the near future.

Finally, one of the reasons for doing the present study was that identical consonants might be useful for morphological decomposition. As a preliminary attempt to answer the question how useful identical consonants are I counted the number of morpheme boundaries that were marked by identical consonants in a short Swedish text. The following text was selected from the introduction of a writing style guide (*Svenska Skrivregler*, Utgivna av Svenska Språknämnden, 1991):

Var-för behöv-er man skriv-regl-er. Tal är ett språk för öra-t, skrift är ett språk för öga-t. I tal-et ha-r vi fler-a sätt att signal-era hur det vi säg-er skall upp-fatta-s: ton-fall, paus-er, beton-ing, röst-styrka, och dess-utom kan vi an-vänd-a gest-er och min-er. I skrift måste vi an-vänd-a andr-a signal-er, som stycke-in-del-ning, stor eller lit-en bok-stav, skilj-e-tecken. De o-lik-a signal-er-na i tal-språk lär vi oss huvud-sak-lig-en spontan-t de är natur-lig-t fram-vuxn-a, men för skrift-språk-et behöv-s viss-a bestäm-d-a regl-er för den yttre ut-form-ning-en.

In this text, there are 56 morpheme boundaries (marked by dashes) and 79 word boundaries (marked by spaces). Out of these 135 boundaries, 26 (19.3%) were marked by identical consonants. On the contrary, there was only one place were two identical consonants did not mark a boundary, namely within the word *spontan-t* (spontaneously). This suggests that the presence of two identical consonants is a limited but rather reliable source of information that is useful for the identification of morphological boundaries.

5. References

Clements G, Hume E 1995 The internal organization of speech sounds. In Goldsmith J (ed), *The handbook of phonological theory*. Oxford, Basil Blackwell Ltd. pp 245–306.

Cutler A, Norris D 1988 The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance* 14: 113–121.

Elert C 1989 *Allmän och svensk fonetik* [General and Swedish phonetics]. Stockholm, Norstedts Förlag AB.

Hedelin P 1997 *Norstedts svenska uttalslexikon* [Norstedts Swedish pronunciation dictionary]. Svenska Förlag: Norstedts Ordbok.

Hock H, Joseph B 1996 *Language history, language change, and language relationship.* New York, Mouton de Gruyter.

McCarthy J 1986 OCP effects: gemination and antigemination. *Linguistic Inquiry* 17: 207–263.

McQueen J 1998 Segmentation of continuous speech using phonotactics. *Journal of Memory and Language* 39: 21–46.

Sigurd B 1965 Phonotactic structures in Swedish. Lund, Berlingska Boktryckeriet.

Smith N 1973 *The acquisition of phonology*. Cambridge, Cambridge University Press.

Sproat R 1992 Morphology and computation. Cambridge MA, The MIT Press.

Stemberger J 1981 Morphological haplology. *Language* 57: 791-817.

Vihman M 1978 *Consonant harmony: its scope and function in child language*. In Greenberg J (ed) *Universals of language, vol. 2, phonology*. Stanford CA, Stanford University Press, pp 281–334.