WHAT MAKES WORDS SOUND SIMILAR IN SPANISH?

Mónica Tamariz-Martel Mirêlis

TAAL, The University of Edinburgh

monica@ling.ed.ac.uk

ABSTRACT

The experiment presented in this paper attempts to establish what parameters are important in the perception of phonological similarity between words. The results broadly support those obtained with a computational approach. We analysed the importance of single segments, vowels vs. consonants, syllabic structure and stress. We also discuss the implications for the role of morphology in the perception of word-form.

1 INTRODUCTION

The aim of this paper is to explore the relative importance of different parameters of word-form similarity such as sharing the same vowels, consonants, or stress.

In last year's post-graduate conference (Tamariz, 2002), we presented a computational approach to this question, and obtained a ranked list of parameters that seemed to indicate that word forms are similar when they share morphological and phonological factors. That approach was based on the finding of a significant correlation between word form and word meaning in English: to a small, but statistically significant extent, words that sound similar tend to have similar meanings. This correlation could help word acquisition in children and new word comprehension in adults. We applied the same measure of semantic similarity to a Spanish corpus, and run a "hill-climbing" algorithm that maximized that useful correlation to return values for the different parameters of word-form similarity.

This computacional approach needed some psychological backing, such as an experiment that obtained different relative values for the same parameters of word-form similarity, but this time judged by people instead of returned by a series of operations performed on corpus data.

2 THE EXPERIMENT

This experiment was carried out on the internet. Participants were recruited through a message sent to a linguistics web forum and also to some friends requesting them to take

part in an experiment and forward the message on to their acquaintances. The note directed participants to a web form containing the instructions and the experimental material. At the end of the form there was a small questionnaire where they were asked about their region of origin, age group, sex and the strategy they had followed while doing the experiment (simply looking at the words, reading them in their heads or reading them out loud).

2.1 Participants

All participants had Spanish as their mother tongue and lived in Spain, in a Spanish-speaking environment (e.g. not in Cataluña, where Catalan is spoken by a large proportion of the population. 23 participants from 7 different Spanish regions took part in this on-line experiment. Nine were male and 14 female. One participant was between the ages of 10 and 19, seven between 20 and 29, nine between 30 and 39 and six between 40 and 49. Fourteen participants reported having read the words out loud, eight to have read them in their heads, and only one to have simply looked at them.

2.2 Materials

2.2.1 Parameters

The parameters are features that two words can have in common. We used two groups of stimuli, cv-cv and cvccv non-words. Table 1 shows the parameters used in this experiment for each of the stimulus groups.

cv-cv	**cvccv**
Same 1st consonant	Same 1st consonant
Same 2nd consonant	Same 2nd consonant
Same 1st vowel	Same 3rd consonant
Same 2nd vowel	Same 1st vowel
Same two syllable-initial consonants	Same 2nd vowel
Same two vowels	Same two syllable-initial consonants
Same stress (on 1st syllable)	Same consonant cluster
Same stress (on 2nd syllable)	Same two vowels
Same stressed vowel in the 1st syllable	Same stress (on 1st syllable)
Same stressed vowel in the 2nd syllable	Same stress (on 2nd syllable)
	Same stressed vowel in the 1st syllable
	Same stressed vowel in the 2nd syllable
	Same syllabic structure (*cvc-cv* or *cv-ccv*)

Table 1. Parameters used in the experiment for cv-cv and cvccv words.

2.2.2 Stimuli

We prepared a set of 93 triads (like the one shown in Figure 1) that represented all the sensible combinations of parameters for four and for five phoneme non-words (cv-cv and cvccv, respectively). E.g. in the example triad in Figure 1, *mélto* shares the third consonant with the base non-word *súnta* and *múlko*, the stressed vowel on the first

syllable, so this the triad compares parameters "same 3rd consonant" vs. "same stressed vowel on the 1st syllable".

o mélto
súnta
o múlko

Figure 1. One example of non-word triad. In this case the top word on the right shares the third consonant (t) with the word on the left and the bottom word shares the stressed vowel in the first syllable (ú). These are the two parameters that we are comparing here.

Since every words must be stressed on one syllable, when stress was not an issue all three words in a triad would share the same stressed syllable.

All the possible and sensible combinations of parameters were used. Parameter combinations that were impossible to occur simultaneously such as "sharing the stress on the first syllable vs. sharing the stress on the second syllable" were excluded. Also, in order to keep the number of stimuli to a minimum, we excluded combinations that would generate an obvious response, e.g. "sharing the vowel 1" vs. "sharing vowels 1 and 2". Here we assumed that the second option would be rated as more similar than the first one and for the results, gave it a confidence factor of 0.75 (see Table 3 below). The full set of triads is found in Appendix 1.

2.2.2.1 Frequency

In order to make the non-words natural to the Spanish ear, the frequencies of the consonants in the sets of non-words mirrored the frequencies of consonants in the word sets extracted from the corpus and used in the hill-climbing algorithm. For cv-cv words, the similarity between the distribution of the first and the second consonants was highly significant (t-test, $p < 0.001$). For cvccv words, the similarity of consonant clusters was significant (t-test, $p < 0.003$) but the similarity of first consonants was not (t-test, $p < 0.09$). Note that given the constrained set of consonant clusters in cvccv words in Spanish, there are not many phoneme combinations that are not real words.

2.2.2.2 Neighbourhood

The phonological neighborhood density (number of words that sound similar to a target word) of the stimuli was calculated using a 707,000 word (including derived and inflected words) corpus of spontaneous speech (UAM corpus, Marcos Marín, 1992). We counted as neighbours: (a) words of the same length that differed from the stimuli by a 1-phoneme substitution; (b) words up to 6 phonemes (for 4-phoneme stimuli) and up to 8 phonemes (for 5-phoneme stimuli) that contained the stimulus; and (c) longer words whose coda was the stimulus i.e. that rhymed with the stimulus. E.g. the stimulus non-word *síto* has 27 neighbours:

Differs by 1 phoneme from (4-phoneme neighbors): Cíto, Ríto, kíto, míto, píto, sEto, sído, sígo, síko, síno, site, sítu, sOto, títo, zíto.

Is contained by (5- or 6-phoneme neighbors): osito, ositos, pasíto, písíto, besíto.

Rhymes with (7 or more phoneme neighbors): Rekísito, Repasito, bersito, deskansito, ekskísito, nezesito, konkursito, luísito.

	cv-cv	***cvccv***
Neighbourhood range	0-77	0-45
Average	8.2	0.7
Distr. kurtosis	23.4	85.3

Table 2. Data about the neighbourhood density of cv-cv and cvccv stimuli.

Table 2 shows that cv-cv words have larger neighbourhoods, and that they are more evenly distributed. For the complete neighbourhood density list, see Appendix 2.

2.3 Method

Each participant saw 45 randomly selected triads so as to keep the experiment time low and encourage participation and completion. Participants were asked to read the non-word triads and determine which of the two words on the right *sounded* more similar to the word on the left. It could be argued that in an experiment concentrating on phonological and morphological aspects of the word-form stimuli should be acoustic, but in order to access Spanish participants living in Spain the experiment would have to be done over the Internet, and we decided that sound-playing equipment and quality in remote terminals would not be reliable. Therefore the instructions stressed the fact that they should focus on the sound of the stimuli. They were also directed to pay attention to the stress of the stimuli, which was marked in all of them by means of an acute on the corresponding vowel (the usual orthographic stress mark in Spanish). The results of each run were automatically emailed back to the experimenter, together with the demographic data.

3 RESULTS

We analyzed the results for 4 and for 5-phoneme words separately. For each pairwise comparison of parameters we counted the proportions of respondents who had chosen each of the two options to obtain the "winner" of that comparison. E.g. for the triad comparing "having the same second consonant" vs. "having the same first vowel", 1/3 of the respondents preferred the consonant and 2/3 preferred the vowel. We then calculated a factor between zero and one that expressed the confidence of the result, such that if everybody prefers the same parameter the confidence factor for the winner is 1 and if the responses were fifty-fifty, the confidence factor is 0, and there is no winner. In our example, 1/3 (=0.33) more people preferred the winner (the consonant) than the loser (the vowel), so for this comparison we would have winner = second consonant; confidence factor = 0.33. These results are shown in Table 3.

	c1		c2		v1		v2		Tc		Tv		a1		a2	
c1	***W***	***cf***														
c2	c1	1.0	***W***	***cf***												
v1	c1	.33	c2	.45	***W***	***cf***										
v2	c1	.4	v2	.33	v1	.07	***W***	***cf***								
tc	**tc**	**.75**	**tc**	**.75**	tc	1.0	tc	.6	***W***	***cf***						
tv	tv	.6	tv	1.0	**tv**	**.75**	**tv**	**.75**	---	0	***W***	***cf***				
a1	a1	.16	a1	1.0	a1	.09	a1	.8	---	0	tv	.16	***W***	***cf***		
a2	a2	.28	a2	.6	v1	.27	a2	.27	---	0	tv	.27	**---**	**0**	***W***	***cf***
av1	---	0	---	0	**av1**	**.75**	av1	.5	tc	.66	**---**	**0**	av1	.75	**---**	**0**
av2	av2	.46	av2	.55	av2	**.55**	**av2**	**.75**	av2	.33	**---**	**0**	**---**	**0**	av2	1.0

Table 3. Matrix of the winner and the confidence factor for each pairwise combination of parameters. W = winner. Cf = confidence factor. (In bold, not-tested, assumed values.)

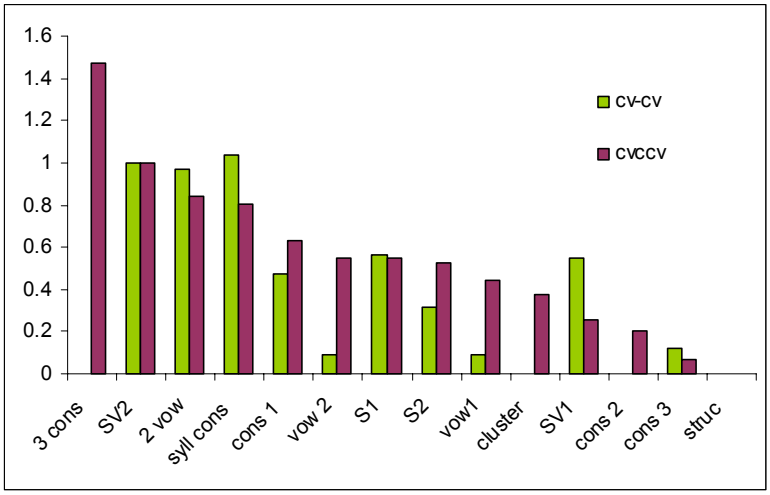

Figure 2. Parameter values obtained for cv-cv and cvccv words.

Then for each parameter we calculated the sum of the confidence factors of the times it had been the winner. These results (normalized) for cv-cv and cvccv words can be seen in Figure 2. A preliminary glance at the results tells us that they make sense, e.g. sharing three or two consonants has a higher value than sharing one consonants, and sharing two vowels, higher than only one. Also, we see a consistency across word-groups: The values of the parameters common to cv-cv and cvccv groups are significantly correlated (R^2=0.57, $p<0.001$). (In the hill-climbing algorithm this correlation is R^2=0.48, $p<0.01$).

3.1.1 Single segments

Figure 3 shows a representation of the relationships between different single-segment parameters (single consonants and vowels). The arrows go from the winner to the loser in each pairwise comparison and the thickness of the lines reflects the confidence factors (also shown beside each arrow).

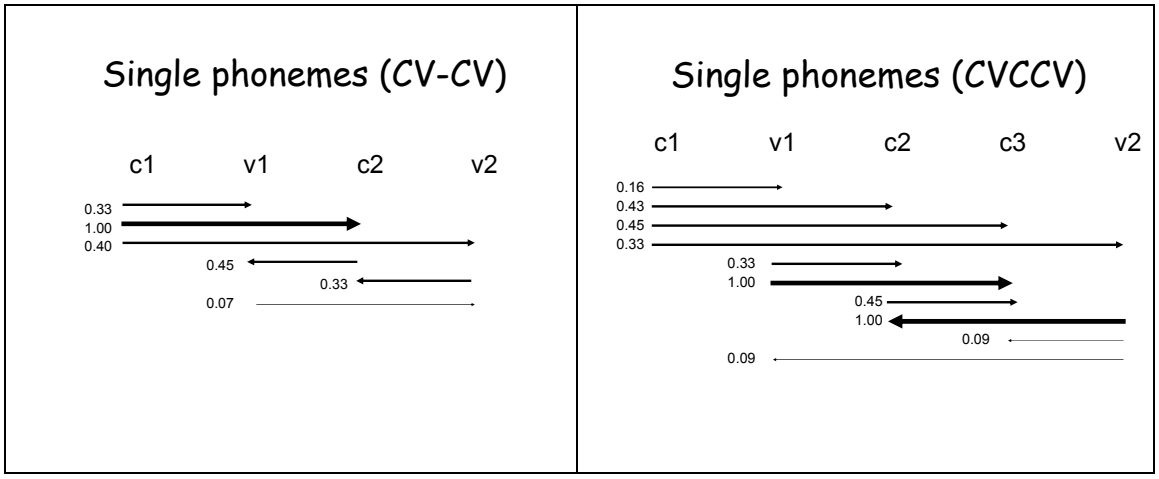

Figure 3. Relationships between the single segment parameter values in cv-cv and cvccv word groups.

The first consonant wins over every other parameter in both groups, and most of the arrows point towards the end of the word. These two facts indicate that people focus more on the beginning of the word to try to find similarities and differences between words, that is, to identify the word.

3.1.2 Vowels vs. consonants

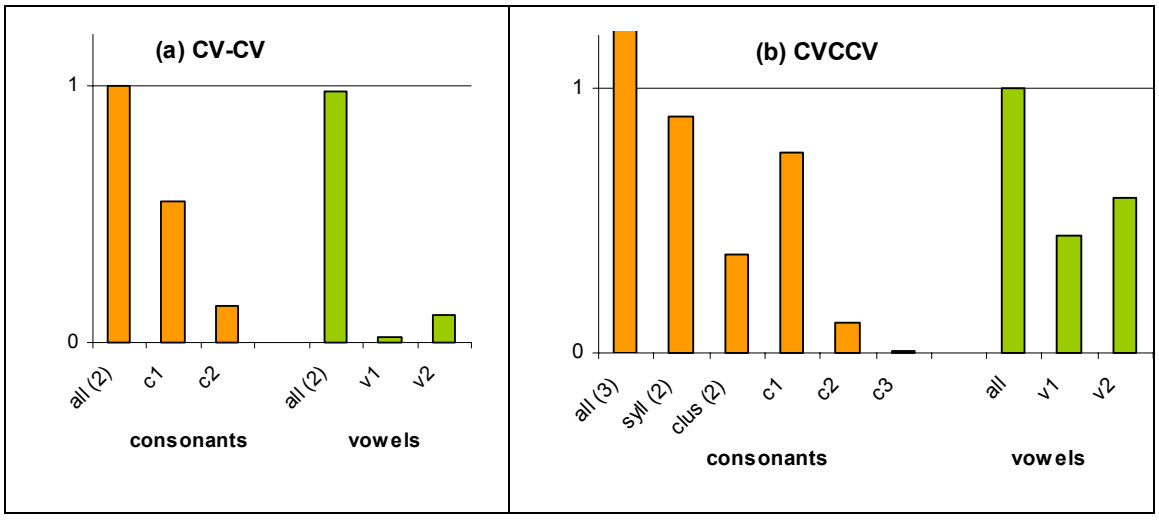

Figure 4. Values obtained by the parameters related to consonants and vowels (a) in cv-cv and (b) in cvccv words.

As seen in Figure 4(a), for cv-cv words single consonants are more important than single vowels for the perception of word-form similarity. However, sharing all the vowels is as important as sharing two consonants. Figure 4(b) shows that for cvccv words, sharing all (two) vowels scores lower than sharing all (three) consonants, but higher than sharing two consonants. This implies that the vocalic structure of the word is almost as important as the consonant structure for the perception of word-form similarity.

3.1.3 Syllabic structure

This parameter only applies to cvccv words, and compares two possible syllable segmentations: cv-ccv vs. cvc-cv, e.g. *mer-ta* vs. *me-tra*. This parameter lost to every other parameter, so we can say that it is of little importance to the perception of word-form similarity.

3.1.4 Stress

The experiment included four parameters related to stress: same stress on the first or second syllable, and same stressed vowel on the first or second syllable. We found that sharing the same stressed vowel on the second syllable was the winner over every other parameter, including sharing all three consonants in the cvccv group. It is worth noting here that the vast majority of two-syllable words in Spanish are stressed on the first syllable, so stress on the second syllable is a marked feature.

3.1.4.1 Morphology

This parameter was also at the top of the ranking obtained with the computational study mentioned above. In that study, almost all the words stressed on the second syllable were verbs, and the stressed vowel (*á, é, í* and *ó*) was the morpheme indicating tense and person, so we hypothesized that the algorithm was giving morphology a role in word-form similarity.

The present experiment the stimuli were non-words. However, we cannot claim that the use of non-words precludes the perception of word-final phonemes as morphemes. E.g. the non-word *bunkí* could be perceived as the 1st person singular of the past tense of the non-verbs bunker or bunker. If morphology perception interferes with phonology perception, in the triad [*bunkí* (*teská* or *tesmí*)], *tesmí* could be found more similar to *bunkí* because it could be perceived to be sharing the same tense and person.

In an attempt overcome this problem, we included stimuli ending in *ú*, which is not a verbal morpheme. However, in such triads participants still found words sharing the stressed ú more similar than those sharing any other parameter, including sharing the three consonants. E.g. all participants responding to triad [*kandú* (*kindá* or *pirgú*)] found *pirgú* was more similar to the base word. Morphology, then, cannot be directly responsible for the high score of the parameter "same stressed vowel on the second syllable", but the fact that important information such as segmentation cues and morphology occurs at word-ends, people tend to focus on any phonological variation there, particularly on marked features.

4 CONCLUSION

This preliminary experiment to determine what parameters are important in making words sound similar broadly supports the results obtained with an algorithm that maximizes the correlation between word form and meaning in a corpus.

When trying to find word-form similarity, people focus more on the beginning of the word, particularly on the first segment; the vocalic structure is almost as important as the

consonant structure; and stress on the second syllable is most important, perhaps because it is correlated with morphology encoding.

ACKNOWLEDGEMENTS

This research has benefited from the support of EPSRC studentship award nr. 00304518.

BIBLIOGRAPHY

Marcos Marín, F. (1992). Corpus oral de referencia del español, Madrid: UAM.

Tamariz, M. (2002) Parameters of word-form similarity in Spanish. Proceedings of the 2002 Postgraduate Conference. Theoretical and Applied Linguistics, The University of Edinburgh. http://www.ling.ed.ac.uk/~pgc/archive/2002/proc02/tamariz02.pdf

Appendix 1. Stimulus sets for cvcv and cvccv words. Words 2 and 3 each share one of the phonological similarity parameters indicated in column one with word 1.

param.	*let.*	*word 1*	*Word 2*	*word 3*
c1-c2	5	búnta	Bísko	línko
c1-c3	5	káste	Kíndo	bínto
c1-v1	5	sárke	Sónti	pánti
c1-v2	5	mínde	Mórka	kórke
c1-tc23	5	rásli	Rónte	bósle
c1-tv	5	fínto	Fáste	kísto
c1-str	5	bésto	Búgra	túnka
c1-a1	5	kórpa	Kengú	méngu
c1-a2	5	sultó	Sánde	pandé
c1-av1	5	tárbo	Túnte	kánte
c1-av2	5	kurtá	Kombé	sondá
c2-c3	5	lórdi	Pérku	péndu
c2-v1	5	linká	Bentó	bistó
c2-v2	5	gúsmi	Tésba	térbi
c2-tc13	5	mórfa	Serpo	melfo
c2-tv	5	pósti	Tésto	tórti
c2-str	5	dákme	Mógri	mónsi
c2-a1	5	bésta	Tusgó	túlgo
c2-a2	5	tuská	Nósde	nordé
c2-av1	5	mólka	Gálpe	góspe
c2-av2	5	pustó	Leská	lenkó
c3-v1	5	pórda	Mésdi	mósti
c3-v2	5	társe	Bínde	bínso
c3-tv	5	ménto	Sarti	sérmo
c3-str	5	bísle	Dáblo	dángo
c3-a1	5	lúmpe	Jospá	jósta
c3-a2	5	bundó	Tálde	talpé
c3-av1	5	súnta	Mélto	múlko
c3-av2	5	tonké	Perká	perté
v1-v2	5	párti	Lánde	lóndi
v1-tc13	5	tíngu	sírka	tórga
v1-tc23	5	rósta	bónde	búste
v1-str	5	gánti	mágle	móske
v1-a1	5	tílpa	kindá	kúnda
v1-a2	5	pírkó	tínka	tenká
v1-av2	5	sinká	mistó	mestá
v2-tc13	5	tónse	lúrde	túrsa
v2-tc23	5	sáldi	pérbi	példo
v2-str	5	múlde	kábre	kánfo
v2-a1	5	sórga	mendá	méndi
v2-a2	5	bóndé	tálke	talkí
v2-av1	5	tónde	rúspe	róspa
tc13-tc23	5	lésta	lónti	kósti
tc13-tv	5	bísna	búlne	tílka
tc13-str	5	minle	maklo	dárso
tc13-a1	5	férna	falnó	páldo
tc13-a2	5	jéntó	júlta	pulká
tc13-av1	5	bárke	búnko	gánto
tc13-av2	5	rendá	risdó	tisbá
tc23-tv	5	túrke	mórka	múnze
tc23-a1	5	pánte	luntí	lúsdi
tc23-a2	5	fústó	mésta	melgá
tc23-av1	5	méspa	bíspo	bérto
tc23-av2	5	pulká	golké	gorbá
tv-str	5	kónda	bósta	bútre
tv-a1	5	góspi	toldí	tálde
tv-a2	5	rándé	tárge	torgú
str-a1	5	mésda	portí	pótri
str-a2	5	tínká	púrde	pugré
str-av1	5	kéndo	mírga	mégra
str-av2	5	fasté	turpó	túblé
a1-av1	5	pésta	dúrko	dérko
a2-av2	5	kustó	perká	perkó
c1-c2	4	kátu	kóbe	róte
c1-v1	4	sípo	sáne	kíne
c1-v2	4	máke	míto	líte
c1-tv	4	pína	pébo	tíba
c1-a1	4	lóga	lasé	máse
c1-a2	4	pité	púro	kuró
c1-av1	4	dúka	dóse	lúse
c1-av2	4	letí	lomé	bomí
c2-v1	4	lóri	péru	póku
c2-v2	4	kábu	díbe	dípu
c2-tv	4	bóra	kíre	kóna
c2-a1	4	síre	maró	mádo
c2-a2	4	bagú	rígo	risó
c2-av1	4	lúko	dáke	dúre
c2-av2	4	daké	pokí	poré
v1-v2	4	súla	múte	míle
v1-tc	4	zúki	púna	zóka
v1-a1	4	kéla	bedó	bído
v1-a2	4	tiká	piré	poré
v1-av2	4	masó	palé	puló
v2-tc	4	búse	táre	báso
v2-a1	4	táro	buló	búle
v2-a2	4	dolú	séru	serí
v2-av1	4	mále	róse	rási
tc-tv	4	kúte	káto	dúbe
tc-a1	4	káli	keló	péjo
tc-a2	4	puné	póna	kodá
tc-av1	4	síto	sáte	míle
tc-av2	4	milá	molé	botá
tv-a1	4	néko	tejó	túja
tv-a2	4	kasí	dári	deró
a1-av1	4	séli	túka	téka
a2-av2	4	siró	kaní	kanó

code	*parameter*
c1	same 1st consonant
c2	same 2nd consonant
c3	same 3rd consonant
v1	same 1st vowel
v2	same 2nd vowel
tc13	same 1st and 3rd consonants
tc23	same 2nd and 3rd consonants
tv	same two vowels
str	same syllabic structure
a1	same stress (1st syllable)
a2	same stress (2nd syllable)
av1	same stressed vowel (1st syllable)
av2	same stressed vowel (2nd syllable)

Appendix 2. Phonological neighbourhood density for the cv-cv and cvccv stimuli.

CV-CV Stimuli	
Word	***Dens.***
tIba	77
mAdo	41
bIdo	33
sIto	27
pIna	25
kAto	24
mIto	23
kOna	19
tEka	19
pOna	18
bAso	16
pEbo	15
bOra	14
kEla	13
tAro	13
mAse	13
lOga	12
mAle	12
pUro	12
sUla	11
kAli	11
kObe	11
zOka	11
mAke	10
mIle	10
mIle	10
lUko	9
dOse	9
kIre	9
sIpo	8
dUka	8
sEli	8
dIbe	8
marO	8
rIgo	8
pUna	8
tAre	8
sAte	8
sIre	7
kasI	7
sAne	7
dAke	7
tUka	7
kIne	7
dUbe	7
pEjo	7
lOri	6
kUte	6
rOte	6
kanO	6
masO	5
pirE	5
palE	5
kelO	5
dAri	5
dUre	5
botA	5
pitE	4
bUse	4
nEko	4
pEru	4
mUte	4
rOse	4
molE	4
lIte	4
lUse	4
porE	4
porE	4
serI	4
rAsi	4
kAbu	3
dakE	3
sirO	3
pokI	3
kanI	3
pOku	3
bUle	3
tUja	3
derO	3
kAtu	2
milA	2
lasE	2
lomE	2
bedO	2
sEru	2
risO	2
kodA	2
letI	1
bagU	1
zUki	1
punE	1
kurO	1
tikA	0
dolU	0
bulO	0
tejO	0
bomI	0
dIpu	0
pulO	0

CVCCV Stimuli	
Word	**Dens.**
tArse	45
mEnto	40
kAnte	10
bErto	4
kAste	3
pAnte	3
pEsta	2
bentO	2
tEsto	2
bIspo	2
tOrga	2
sultO	1
mOrfa	1
sUnta	1
lEsta	1
fErna	1
bArke	1
kOnda	1
lAnde	1
lUrde	1
kAbre	1
mendA	1
portI	1
pEndu	1
bistO	1
dAngo	1
pAldo	1
perkO	1
All the rest	0