Frequency Distributions for Names and Unconstrained Words Associated with the Letters of the English Alphabet

James R. Lewis

IBM Pervasive Computing Division
8051 Congress Ave, Suite 2227
Boca Raton, FL 33487
jimlewis@us.ibm.com

Abstract

This paper provides frequency distributions (constrained and unconstrained) for words and names associated with the letters of the English alphabet. These distributions can be useful when developing voice spelling user interfaces (Lewis & Commarford, 2003). The data suggest that the unconstrained word distribution is preferable to the names distribution because, contrary to expectation, the overall consistency of unconstrained responses was not poorer than the responses in the names distribution and the occurrence of cases in which participants did not provide a response for the names distribution was significantly greater than for the unconstrained distribution.

1 Introduction

Voice spelling is an important component of many speech systems (embedded in other appliances, accessed by telephone, or for hands-free use of a desktop dictation system). If a user needs to use a word that is not in the system's built-in lexicon and either cannot or chooses not to use a keyboard, then the user must spell the word by voice.

The standard approach to voice spelling is to let users say the letter names or, because some letter names have high acoustic confusability, to also permit the use of the military alphabet (alpha, beta, Charlie, etc.). The military alphabet has an advantage over letter names in that the words chosen for the military alphabet have very low acoustic confusability. The military alphabet has a disadvantage over letter names in that it is somewhat difficult to memorize the military alphabet.

The purpose of the current research was to find out what words come first to people's minds when asked to produce a word for each letter of the alphabet. In one experiment participants provided words without constraint. In a separate experiment, participants provided first names for each letter. The reason for constraining participants in this way was to see if participants under this condition produced more consistent responses than those in the unconstrained condition.

2 Method

2.1 Participants

The source for participants in each experiment was a set of 400 IBM employees (800 employees in all, 400 for each experiment), selected at random from an internal e-mail directory of all of the IBM employees in the United States. Each participant received an e-mail invitation to participate. Of the employees invited to participate, 103 (26%) responded to the invitation for the first experiment (no constraints) and 120 (30%) responded to the invitation for the second experiment (names for letters).

2.2 Materials

I used WebSurveyor[1] to construct a web-based form for the evaluation. The form simply provided a space by each letter of the alphabet (arranged vertically on the page in alphabetical order), with instructions appropriate for the specific experiment (to either provide any word for each letter, or to provide only first names).

2.3 Procedure

After receiving the e-mailed invitation, participants clicked a link in the message that brought up the WebSurveyor page containing the survey. Participants read an introduction explaining the purpose of the survey, then completed and submitted the form. The WebSurveyor program kept track of the raw data. After collection, I used a spreadsheet to organize, summarize, and analyze the results.

3 Results

From the survey data I created a set of 52 tables, one for each letter of the alphabet by production condition (unconstrained or constrained). Each table contained the submitted words, listed in decreasing frequency of submission. Each table also showed the cumulative frequency and indicated (1) the point at which the cumulative frequency for a given frequency of submission came as close as possible to 80%, (2) the number of words included in the list up to that point, (3) the standard deviation of the frequencies, and (4) the total number of different words submitted. The choice of 80% as a cutoff point was arbitrary, but seemed reasonable for the purpose of ensuring fairly comprehensive coverage and for comparing the cutoff points of the different distributions. See the Appendix for a summary table of the three most-frequently produced words for each condition (for the complete set of tables, see Lewis, 2001).

I computed the correlations among the 80% cutoff points, standard deviations, and total word counts for both word distributions. In both distributions the cutoff points and total word counts had a positive correlation ($r = .84$, $p = 0.0000001$ for unconstrained; $r = .50$, $p = 0.01$ for name-constrained), and both of these variables had negative correlations with the standard deviations (r ranging from -.64 to -.87, all $p < .0004$). These correlations show the interesting (though perhaps not surprising) fact that the more words produced for a letter (which increases the 80% cutoff point and the total word count), the smaller the standard deviation.

[1] WebSurveyor is a trademark of WebSurveyor Corp.

A set of *t*-tests (treating letters as subjects) indicated that there was no significant difference between the distributions for the variables of 80% cutoff point, standard deviation, and total word counts. There was a significant difference between the distributions for the variable of Nones – the percentage of responses for which respondents indicated that they couldn't think of a word for a given letter ($t(25) = 3.61$, $p = .001$). This never occurred in the unconstrained distribution, but happened for a number of letters in the names distribution (most notably, Q, U, X, Y and Z).

4 Discussion

The hypothesis that asking for first names would produce more consistent responses than completely unconstrained responding did not hold. For some letters this was true, but for others it was not. Overall, the distributions had similar statistical properties with regard to 80% cutoff points, standard deviations, and total word counts (all measures of response consistency). However, the number of letters for which participants were not able to think of any appropriate response was greater for first names than for unconstrained words. For this reason, designers should avoid requiring users to provide first names when using words to perform voice spelling. One reasonable approach for limited voice spelling grammars would be to provide three candidates for each letter – the word from the military alphabet, the word with the greatest frequency for each letter from the unconstrained tables, and the word with the greatest frequency for each letter in the first names tables. Depending on the capacity of the system, it would be reasonable to include more of the words from each source to increase the odds of having a match to a user's choice of word when voice spelling.

Another use for the results of these experiments is in the spoken feedback of alphanumeric information. Many current systems simply feed back letter names, but the confusability of letter names is a liability in spoken language output as well as in spoken language input. Some systems use the military alphabet for feedback (for example, "A as in alpha"). The output of such a system would sound more familiar (less formal) if it used more familiar words, such as "A as in apple".

5 References

Lewis, J. R. (2001). *Frequency distributions for names and unconstrained words associated with the letters of the English alphabet* (Tech. Report 29.3437). West Palm Beach, FL: International Business Machines Corp. (http://drjim.0catch.com/vspellwords.pdf)

Lewis, J. R., and Commarford, P. M. (2003). Developing a voicespelling alphabet for PDAs. *IBM Systems Journal*, *42*, 624-638.

6 Appendix. The Three Most-Frequently Produced Words as a Function of Constraint and Letter

Word (Unconstrained)	Percent	Cumulative	Word (Name)	Percent	Cumulative
apple	67.0	67.0	Adam	10.8	10.8
alpha	11.7	78.6	Alan	9.2	20.0
able	4.9	83.5	Albert	5.8	25.8
boy	40.4	40.4	Bob	33.3	33.3
baker	9.6	50.0	Brian	8.3	41.7
bravo	9.6	59.6	Barbara	6.6	48.3
cat	38.5	38.5	Charlie	32.3	32.3
Charlie	37.5	76.0	Cathy	12.8	45.1
Charles	2.9	78.8	Charles	7.5	52.6
dog	66.0	66.0	David	45.8	45.8
David	13.6	79.6	Dave	9.2	55.0
delta	8.7	88.3	Dog	7.5	62.5
elephant	21.7	21.7	Edward	29.4	29.4
Edward	14.2	35.8	Ed	7.6	37.0
echo	12.3	48.1	Eric	5.0	42.0
Frank	34.0	34.0	Frank	63.8	63.8
fox	21.4	55.3	Fred	11.8	75.6
foxtrot	4.9	60.2	Fox	4.7	80.3
George	25.0	25.0	George	48.8	48.8
girl	16.3	41.3	Gary	9.9	58.7
good	9.6	51.0	Greg	9.9	68.6
help	13.6	13.6	Harry	32.5	32.5
Henry	13.6	27.2	Henry	25.8	58.3
Harry	11.7	38.8	Harold	5.0	63.3
igloo	13.5	13.5	Irene	11.1	11.1
India	12.5	26.0	Ian	10.3	21.4
Indian	7.7	33.7	Isabel	9.4	30.8
Jack	16.5	16.5	Jack	25.6	25.6
jump	11.7	28.2	John	23.1	48.7
John	9.7	37.9	James	6.8	55.6
king	19.6	19.6	Kevin	16.2	16.2
kite	16.7	36.3	Karen	14.3	30.5
kilo	10.8	47.1	Ken	13.3	43.8
Larry	22.3	22.3	Larry	40.0	40.0
love	10.7	33.0	Linda	11.7	51.7
Lima	6.8	39.8	Laura	4.2	55.8
Mary	42.7	42.7	Mary	46.1	46.1
Mike	8.7	51.5	Mike	10.2	56.3
man	6.8	58.3	Michael	9.4	65.6

Word (Unconstrained)	Percent	Cumulative	Word (Name)	Percent	Cumulative
Nancy	49.5	49.5	Nancy	72.9	72.9
nice	5.8	55.3	Nathan	3.4	76.3
no	5.8	61.2	Nick	3.4	79.7
open	21.4	21.4	Oscar	38.1	38.1
Oscar	18.4	39.8	Oliver	6.8	44.9
orange	12.6	52.4	Oprah	6.8	51.7
Paul	23.5	23.5	Paul	40.0	40.0
peter	13.7	37.3	Peter	35.8	75.8
Papa	5.9	43.1	Papa	2.5	78.3
queen	31.1	31.1	Quency	13.6	13.6
quick	11.7	42.7	Quenten	13.6	27.1
Quebec	8.7	51.5	Queen	8.5	35.6
Robert	17.5	17.5	Robert	38.3	38.3
rabbit	7.8	25.2	Roger	9.2	47.5
Romeo	6.8	32.0	Richard	8.3	55.8
Sam	37.5	37.5	Sam	49.2	49.2
snake	3.8	41.3	Sally	8.3	57.5
star	3.8	45.2	Steve	6.7	64.2
Tom	35.3	35.3	Tom	55.0	55.0
tango	6.9	42.2	Thomas	6.7	61.7
Thomas	4.9	47.1	Tony	5.0	66.7
under	20.0	20.0	Ursula	20.5	20.5
uncle	14.7	34.7	Ulysses	10.3	30.8
ugly	8.0	42.7	Uncle	6.0	36.8
vertical	49.0	49.0	Victor	58.3	58.3
victory	16.7	65.7	Vicky	6.7	65.0
violin	4.9	70.6	Victoria	4.2	69.2
water	9.0	9.0	William	42.0	42.0
William	9.0	18.0	Walter	13.4	55.5
whiskey	8.0	26.0	Wayne	5.9	61.3
x-ray	64.6	64.6	Xavier	31.1	31.1
xylophone	16.7	81.3	X-ray	13.4	44.5
Xerox	11.5	92.7	Xena	10.1	54.6
yellow	45.5	45.5	Yolanda	30.5	30.5
yes	13.1	58.6	Yello	6.8	37.3
Yankee	10.1	68.7	Yvonne	5.9	43.2
zebra	75.2	75.2	Zebra	18.4	18.4
zoo	5.9	81.2	Zach	9.6	28.1
zero	5.0	86.1	Zachary	7.9	36.0