Familiarity and Pronounceability of Nouns and Names: The Purdue Proper Name Database

Aimée M. Surprenant[1], Susan L. Hura[2], Mary P. Harper[3], Leah H. Jamieson[3], Glenis Long[4], Scott M. Thede[3], Ayasakanta Rout[4], Tsung-Hsiang Hsueh[3], Stephen A. Hockema[3], Michael T. Johnson[3], John B. Laflan[3], Pramila Srinivasan[3], & Christopher White[3]

[1]Purdue University, Department of Psychological Sciences, West Lafayette, IN 47907; [2]Lucent Technologies, 101 Crawfords Corner Road, Holmdel, NJ 07733; Purdue University, [3]School of Electrical and Computer Engineering and [4]Department of Audiology and Speech Sciences, West Lafayette, IN 47907

Abstract: Ratings of familiarity and pronounceability were obtained for a sample of 199 names and 199 nouns. Frequency and familiarity were more closely related in the proper name pool than the word pool, although the correlation was modest in both cases. Familiarity and pronounceability were highly related for both names and nouns.

Although word-level models of speech recognition have become very successful in the past several years due to the great increase in computer capacity and processing speed, even the most successful models generally require a great deal of specific training in order to reach high levels of recognition. For items such as proper names, which may number in the tens of thousands and have multiple pronunciations, it is impractical, if not impossible, to train on the entire set. Name recognition is of considerable practical interest given the possibilities of building acoustic interfaces to telephone directories or library catalogs, and shows promise as an area of great overlap between human and machine word recognition. Names occupy a unique position in lexical access: they have no inherent meaning thus they require phonological (sound-based) recognition; on the other hand, proper names have aspects of frequency and familiarity that may allow them to act like words already in the lexicon. A promising immediate approach to designing a name recognition system is to incorporate statistical aspects of proper names (frequency and familiarity) directly. There exists relatively little data on the distribution of proper names in the language (see however (1)), and there are a number of reasons to suppose that words and names will be rated differently on frequency and/or familiarity. We expect that those names that are familiar will also be easy to pronounce and, importantly for the computational aspect, ones that will lead to relatively little variability in pronunciation. Less familiar names may be more difficult to pronounce and result in more varied pronunciations. The data below are a first effort at obtaining reliable familiarity ratings for a random sample of surnames.

EXPERIMENTS AND DISCUSSION

In Experiment 1 we collected ratings of the familiarity of words and names. Two hundred surnames were selected at random from the Purdue University phone book and 200 common nouns were chosen at random from the Penn-Treebank corpus (2, 3); one item from each set was discarded due to an error, leaving 199 names (mean frequency = 3.7, SD = 10.6). and 199 nouns (mean frequency = 2.0, SD = 3.5). The distribution of nouns is more normal and somewhat less skewed than the names. Seventy-five Purdue University students participated in the study for course credit; data from 63 subjects who were native monolingual speakers of English were included for analysis. Nouns and names were presented in blocks, with block order counterbalanced between sessions. Each stimulus was rated on a scale of 1-7, with 1 meaning *not familiar* and 7 indicating *very familiar*.

Table 1 lists the distributions of familiarity ratings for the names compared to the nouns. Note that nouns are on the whole much more familiar than names. The modal response for names was 1 (*not familiar at all*), mean = 2.33, median = 1.47, SD = 1.08. The mode for nouns was 7 (*very familiar*), mean = 4.9, median = 4.69, SD = .90. The familiarity ratings for nouns were highly correlated with printed familiarity (4), r=.92. The data also show the expected modest correlation (5) between familiarity and frequency (r = .26). The familiarity ratings for names were compared to the frequency of occurrence in the Purdue phonebook; there was a moderate (and reliable, p<.01) correlation between frequency and mean ratings of familiarity (.42) of names and a small, but reliable (p<.01), negative correlation between the number of syllables and the mean familiarity ranking of names (-.27). Proper names are rated on the whole as much less familiar than nouns, and the distribution of ratings is very highly skewed to the bottom of the scale.

TABLE 1. Responses in each familiarity category

	Familiarity Category						
	1	2	3	4	5	6	7
Names	5922	1904	1209	841	668	496	558
Nouns	462	966	1394	1687	2046	2026	3015

TABLE 2. Responses in each pronounceability category

	Pronounceability Category						
	1	2	3	4	5	6	7
Names	440	533	782	1021	1219	1248	3508
Nouns	7	26	119	309	642	992	6659

In Experiment 2, we obtained ratings of pronounceability from a population similar to the one that took part in the first experiment. It is predicted that those names that are less familiar will also be those that are judged more difficult to pronounce. The same word and name pools described in Experiment 1 were used in Experiment 2. Fifty different students from the same pool rated nouns and names on pronounceability; data from 44 native monolingual speakers of English were included for analysis. Each subject was given 199 words and 199 names to rate on a scale of 1-7, with 1 being *very difficult to pronounce* and 7 being *very easy to pronounce*. Two different random orders of words and names were constructed and order of noun or name rating was counterbalanced across subjects. Subjects were allowed to go at their own pace but were instructed not to go back and change ratings they had already made.

Table 2 shows the distributions of pronounceability ratings for names as compared to nouns. Note that nouns were rated as much easier to pronounce than names, although the difference is not as striking as for familiarity in Experiment 1. The modal response for names was 7, mean = 5.3, median = 6.3, SD = 1.9. The mode for words was also 7, but the mean and medians were higher; 6.6 and 6.8, respectively; SD =.92. There was a correlation of .63 between ratings of pronounceability of nouns and our ratings of familiarity collected for Experiment 1 as well as a substantial correlation (.71) with printed familiarity (4). One other substantial correlation for nouns was a large negative correlation with the number of syllables (-.67). This suggests that the greater the number of syllables in the word, the harder it is to pronounce. Word frequency was only slightly related to pronounceability of nouns. As with the nouns, pronounceability was only marginally correlated with actual frequency (.21; p=.05) in the name ratings. However, the correlation with number of syllables was much smaller than the comparable correlation between syllables and noun pronunciation (-.43). Finally, the correlation between pronounceability ratings and our familiarity ratings was very high (.74).

The data reported here confirm our intuitions that there are substantial differences in how names and nouns are rated. The randomly-selected set of nouns used in these studies were rated as more familiar and easier to pronounce than the randomly-selected set of surnames. The subjective human experience with proper names seems to parallel the increased level of difficulty of automatic speech recognition (ASR) systems dealing with names. Strong correlations were found for both names and nouns between familiarity ratings and pronounceability ratings; subjects rated familiar items as easier to pronounce than unfamiliar items and vice versa. Familiarity was a better predictor of pronounceability than was written frequency for both names and nouns, but frequency counts are much easier to obtain. These data are an important first step in improving performance of computer speech recognition systems on proper names. The results of these studies provide basic distributional information about proper names, and more significantly they describe some ways in which names differ from common nouns. These results will be used to justify, in part, using different processing strategies for names and for incorporating information on familiarity or pronounceability into ASR systems.

ACKNOWLEDGEMENTS

Portions of this work were supported by an Interdisciplinary Research Grant from the Purdue University School of Liberal Arts and Schools of Engineering awarded to A. Surprenant, S. Hura, M. Harper and L. Jamieson.

REFERENCES

1. Zechmeister, E., King, J., Gude, C. & Opera-Nadi, B., *Behavior Research Methods & Instrumentation,* **7**, 531-533, (1975).
2. Marcus, M., Santorinin, B. & Marcinkiewicz, M. A., *Computational Linguistics,* **19(2),** 313-330, (1993).
3. Kucera, H. & Francis, W., *Computational analysis of present-day American English.* Providence, R.I.: Brown University Press, (1967).
4. Coltheart, M,. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology,* **33A**, 497-505, (1981).
5. Gilhooly, K. J. & Logie, R. H., *Behavior Research Methods & Instrumentation,* **12**, 428-450, (1980).