The 'universal' structure of name grammars

And the impact of social engineering on the evolution of natural information systems

Michael Ramscar Asha Halima Smith Melody Dye Richard Futrell Peter Hendrix Harald Baayen Rebecca Starr

University of Tubingen, Stanford University, Indiana University, Massachusetts Institute of Technology, Carnegie Mellon

Abstract

Proper name systems provide individuals with personal identifiers, and convey social and hereditary information. We identify a common information structure in the name grammars of the world's languages, which makes this complex information processing task manageable, and evaluate the impact that the re-engineering of naming practices for legal and political purposes has had on the communicative and psychological properties of these socially evolved systems. While East-Asian naming systems have been largely unaffected by state legislation, legal interference has transformed Western naming practices, making individual names harder to process and remember. Further, the structural collapse of Western naming systems has not affected all parts of society equally: In the US, it has had a disproportionate impact on those sections of society that are least successful in economic and social terms. We consider the implications of these changes for name memory across the lifespan, and for future naming practices.

Keywords: Names; Learning; Memory; Information Theory

What's in a name?

Naming is unique to our species and central to our lives. Names are the primary linguistic means by which we discriminate individuals from their peers, and they are an integral part of our identities. Names also play an important social role: they carry hereditary information that helps regulate marriage between relatives and the transference and distribution of property, as well as fostering group identities that bring cohesion to the conduct of social enterprises such as agriculture, industry, statecraft, and war.

While names for individuals appear to be as old as civilization itself, some functions of names are recent developments (Scott, 1998). Henry VIII decreed that English marital births be recorded under the surname of the father in the 14th Century, but children could, and regularly did adopt different surnames. Hereditary family names only became universal in the UK with the establishment of Her Majesty's Register Office in 1836 (Matthews, 1967). Naming conventions in the Netherlands were formalized by statute in 1811 (Van Poppel, Bloothooft, Gerritzen & Verduin, 1999); in Korea, naming practices were regulated in 1812 (Nahm, 1988), the same year that a Prussian edict granted citizenship to Jews in return for the adoption of fixed patronyms (Scott, Tehranian & Mathias, 2002).

Despite their personal and social importance, names are uniquely difficult to learn and remember (Cohen, & Burke, 1993; Valentine, Brennen & Bredart, 1996). Names produce most naturally occurring tip of the tongue states (TOTs—where one cannot produce a word one is sure one knows) (Rastle & Burke, 1996; Griffin, 2010); patients with cognitive impairments show greater decrements in name retrieval than for other knowledge (Yasuda, Nakamura & Beckman, 2010); and the recall of names appears to be disproportionately impaired in old age. Indeed, many older adults consider deteriorating name memory to be the most disturbing cognitive problem they face (Lovelace & Twohig, 1990).

Here, we identify a common information structure in the name grammars of the world's languages, and reveal the impact that regulating names for legal and political purposes has had on their memorability as populations have grown in the wake of industrialization: while some name systems survive intact, legislation has had a detrimental effect on many name grammars, dramatically undermining their communicative efficiency.

Why names are different—and difficult

While most nouns are generic—spoon, dog, idea—personal names are sui generis: ideally, they uniquely discriminate individuals from their peers. While this could easily be achieved by giving each individual a unique label, this approach would massively increase linguistic complexity. By now, it would have generated a billion extra English words. At points in speech where a name could occur in this kind of a system, entropy—a formal, information theoretic measure of uncertainty (Shannon, 1948)—would vastly exceed anything heretofore encountered. Because entropy peaks accurately predict difficulties in the production and comprehension of speech (McDonald & Shillcock, 2001; Clark & Wasow, 1998), this would cause far more processing problems than actual English names, which are already far more taxing than other vocabulary items.

Thus, while 'one-name-per-person' would eliminate residual uncertainty about the identity of named individuals, it would maximize processing demands in doing so. Realistically, the psychological cost of such a system is too high: Given the highly social, interconnected nature of human life, and our finite information processing capacities, one-name-per-person would prove unworkable as societies developed beyond small groups. Unsurprisingly, no major language has a naming system remotely close to it (Alford, 1987).

A "Universal" Grammar for Names

Although some fine-grained details differ, all the world's major languages have evolved the same solution to the challenges names pose: instead of using unique labels, individual identifiers are formed from hierarchically structured naming tokens. Name grammars enable large sets of identifiers to be constructed out of much smaller sets of

naming words, assigning individuals relatively unique identifiers, while avoiding the outrageous peaks in entropy that would result from a one-name-per-person system.

Family Name	**Clan / Generation Name**	**Given Name**
Least Uncertain	*More Uncertain*	*Most Uncertain*
Baek	Seung	Ki

For example, in the traditional Sinosphere (Matisoff, 1990) naming system used in Chinese-speaking countries and Korea (Kwang-Sook, 2003), names comprise 3 elements: 1) a small number of family names, 2) a clan or generation name (Martin, 2006), and 3) a given name, fairly specific to the individual. (Here, the first parts of physicist Seung Baek Ki's name mean, "a Baek from Suwon." (Kiet, Baek, Jeong & Kim, 2003)). Elements are distributed in these sets in a highly efficient, Zipfian manner (Baek, Kiet & Kim, 2007)2: there are only around 250 Korean family names, three of which are common to around 50% of the population (20% of all Koreans are called Kim). Because of this, names act as efficient hierarchical decision trees: each name element increases the degree to which an individual is identified, while minimizing the entropy at the point each element is encountered (Figure 1).

These distributions have been stable for centuries. As Korea's population grew post-industrialization (Zipf, 1965), the peak entropy of Korean names barely changed. Name grammars in Chinese-speaking countries have developed along similar trajectories.

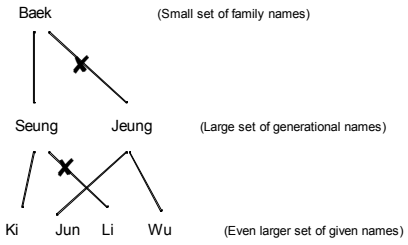


Figure 1. Hierarchical branching minimizes *entropy* (which quantifies the uncertainty produced when a number alternatives that need to be processed at any point) as a Korean name unfolds in time.

This efficient information structure is not unique to Sinosphere names: it is common to the native name grammars of all the world's major languages. For instance, traditionally, English names comprised a given name drawn from a relatively small set, optional middle name(s), and an identifier (in Modern English, a family name), drawn from a large set of personal characteristics, topographic and toponomic features, occupations and patronyms: e.g., John the Farmer, or John White Head. English names rarely contained a specific, unique identifier. Instead, individual identities were the sum of a name's parts: the 11th Century English name in (3) means, "Adam, a farmer from Ramscar" (Ramskir, 1973).

Given Name	**Other Name**	**Identifier**
Least Uncertain	*More Uncertain*	*Most Uncertain*
Adam	Hegger	de Romeskerre

What may be surprising from a modern perspective is that the distributional pattern of names in modern Korea would have been familiar to pre-industrial Europeans: In every 50-year period from 1550 to 1799, around 50% of boys born in England were named William, John or Thomas, and 50% of girls Elizabeth, Mary or Anne (Smith-Bannister, 1997), mirroring the distribution of Korean family names. Given names in other Western and Northern European languages were also distributed this way (Galbi, 2002; Lieberson & Lynn, 2003; Bourin, 1994) even when patronymic conventions were employed (Williams, 1961). Historically, compact, stable Zipfian first name distributions were the norm: prior to industrialization, one-in-five English girls had the first name Mary, just as around one-in-five Korean girls today has the first name Kim. What caused some distributions to change, while others have remained as they were?

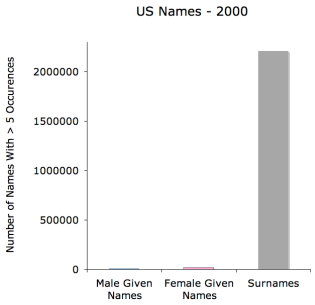


Figure 2. The hierarchical organization of modern English names. The proportion of surnames to given names reflects the fact that given names—which occur before surnames in English—contribute less to individual identities than surnames. The surnames depicted here represent over 95% of the US population in the 2000 census, and the given names over 95% of social security applications in the US in 2000.

The rise of nation states, and their impact on Western name grammars

Ordinarily, nobody knows everybody. However, the development of centralized states created entities that actually did want to know everyone: for the purpose of taxation, conscription, etc. To facilitate this, states regulated names. In the Sinosphere, the burden of coding heredity fell on the first (least diverse) name element. However, in other parts of the world, the final, most diverse element was targeted. In English, idiosyncratic features specific to individuals – such as John the farmer, John with the white head – became fixed hereditary markers, such that bakers might be called Farmer, or redheads Whitehead. English name grammar retained its hierarchical structure (Figure 2). Consequently, as population growth accelerated in the 19th and 20th centuries (Figure 3), two changes occurred:

1. A larger, more diverse set of first names began to be used, increasing the peak signal entropy of names, making them harder to process and recall.

2. More people shared last names, increasing the likelihood of two people having the same name, increasing residual entropy about the individual identified by a name.

Both these changes had an impact on the efficiency of English names.

In the UK in 1801, 3 male names accounted for 52% of male births and 3 female names 53% of female births (the traditional distribution). By 1994, this had dropped to 11% and 9%, respectively. In both instances, the relationship to population growth is strong: r2=.99 & .97. Similarly, while in 1801, 85% of males and 82% of females received 1 of 10 names, by 1994, this had dropped to 28% and 24%; related to population growth, r2=.99 and .97. In 1801, 22% of males and 24% of females in the UK received the most common first name for their sex; in 1994, it was 4% and 3% (both r2=.98). Social Security card applications in the US between 1900-1999 show the same patterns of change in naming: a big decline in the number of babies being given the most frequent names, and an increase in the diversity of given names. (For all reported correlations, $p<0.001$)

*The sample for this survey was taken from the U.S. Social Security Administration (http://www.ssa.gov/oact/babynames/).

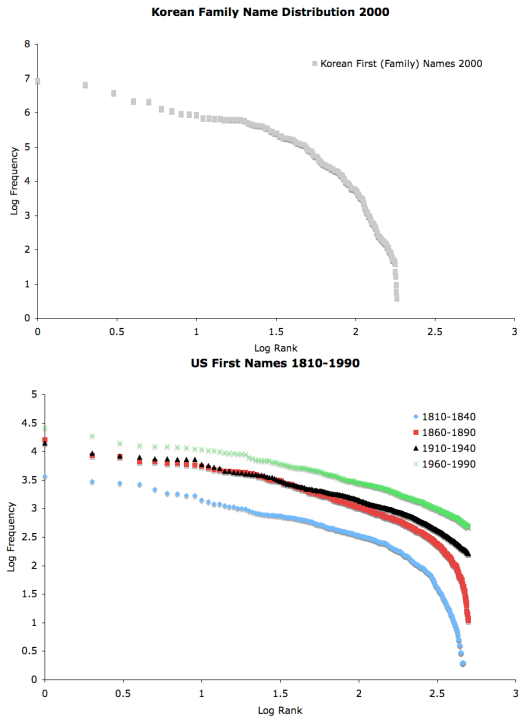


Figure 3. The **top panel** illustrates the distribution by rank frequency of Korean family names in 2000. The **bottom panel** shows the change in the written frequencies of the 500 most common US male names as the United States population grew in the 19th and 20th Centuries (1810-1990). The greater similarity in the distribution of Korean family names and American male names at the beginning of the 19th Century is apparent, as is the change thereafter.

An analysis of the 500 most common male names from 1810-1990 in the Corpus of Historical American English (400,000,000 words) indicates that at the beginning of the 19th century, US given name distributions were similar to Korean family names (Figure 3). While the distribution of Korean family names remained stable, the distribution of American and British given names changed dramatically as populations grew.

Measuring the Effects of Change: A Tale of Two Congresses

Mainland China has experienced massive population growth in the past two centuries, but its traditional Sinosphere naming system has changed little as a result (Yuan, 2002; Mountain, Wang, Du, Yuan, 1992). To illustrate how the changes to the distribution of American names have affected name efficiency, we compared two similarly-sized, naturally-occurring samples of names from each of these two countries: the Senators and Representatives of the 112th United States Congress (n=440), and the members of several subcommittees of the National Committee of the Chinese People's Political Consultative Conference (n=431).

*The samples for this survey were taken from the website of the United States Congress (http://www.house.gov/representatives/) and the website of the Chinese People's Political Consultative Conference (http://www.cppcc.gov.cn/)

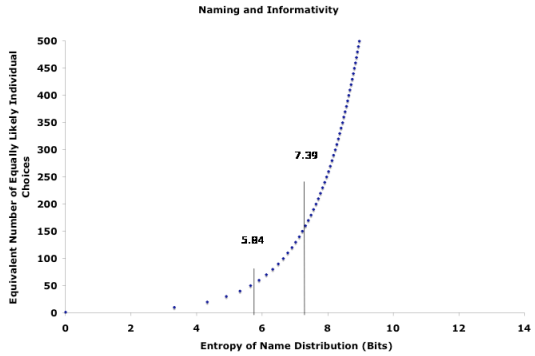


Figure 4. The information challenge posed by names as their numbers grow. The y-axis plots *perplexity*, which describes the entropy (x-axis) in a complex distribution in terms of a number of equally likely alternatives. The perplexity of US names is more than twice that of Chinese names in this sample.

The total entropy of both samples is almost identical (US Congress=8.78 bits; CPPCC=8.75 bits), because in each case, no members share names; accordingly, there is no residual entropy in either sample. There is, however, a marked disparity in the way information is distributed in the samples: the first elements in the CPPCC names contain only 5.84 bits of information (in information processing terms, this is equivalent to differentiating between around 60 equally likely name labels; Figure 4). By contrast, the given (first) names of US Congress members contain 7.39 bits (equivalent to processing around 170 equally likely labels). Both sets of names convey the same amount of overall information, but this information is more evenly distributed in Chinese names: US first name elements impose far greater information processing demands than their Chinese equivalents, and later US name elements are far more redundant (full form analysis–treating Mike / Mick as forms of Michael–reduced US name perplexity to 120,

suggesting that Congressional names would more memorable if nicknames were avoided).

Winners and Losers in the Decline of the US Naming System

Although the efficiency of English names has declined in the past 200 years, the effect of these changes has not been felt in the same way across society. Traditionally, the perplexity of US female names was slightly greater than male names (Figure 5A), but the difference in perplexity between male and female names was relatively constant. However after 1950 this difference began to rise sharply (Figure 5B), increasing the difference in the amount of information that had to be processed in recalling, producing and comprehending female names as compared to male names.

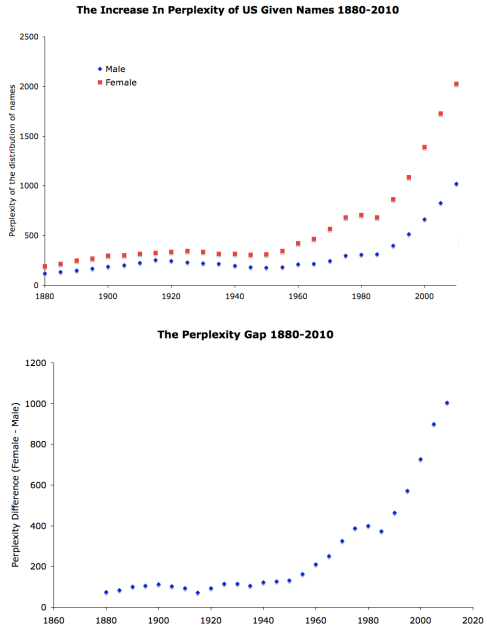
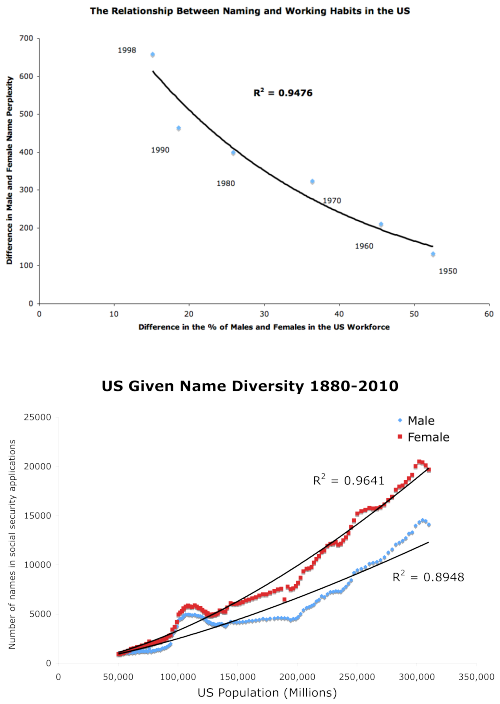


Figure 5. Top: The perplexity of male and female names with a count ≥ 5 in US social security applications at 5 year intervals from 1880 to 2010. **Bottom:** The difference in the perplexity of female names as compared to male names (female perplexity - male perplexity) across this period.

Given that the period since 1950 saw an increase in economic and social equality between males and females in the US (Fullerton, 1999), the close relationship (Figure 6A) between the growth in the perplexity difference between male and female names and the increasing percentage of females in the workforce is surprising, as is the fact that increases in the number of women working outside the home have coincided with an exponential increase in the degree to which female names are harder to process than male names. Figure 6B offers one possible explanation for this: the strong correlation between population size and female name diversity may indicate that parents actually take great care in naming their daughters, but that the constraints that have been imposed on Western naming practices are distorting the intended effect of name choices.

Figure 6. Top: The proportion of men and women over age 16 in the US workforce (male - female percentage) plotted against the different perplexity of female and male names (female perplexity - male perplexity) in the period 1950-1998 (data from the US Bureau of Labor Statistics & US Social Security Administration). **Bottom:** The increase in the US population at five-year intervals from 1880-2010 and the increase in the number of male and female names with a count ≥ 5 in US social security applications.



Nobody Knows My Name

Women are not the only group in the US to have experienced a disproportionate decline in name efficiency. As has often been noted, African American parents systematically choose distinctive given names for their children (Lieberson & Mikelson, 1995). The pattern of this trend is puzzling: although Black parents living in predominantly White communities choose distinctive names for their children, Black parents living in predominantly Black communities choose even more distinctive names. (If it were simply the case that parents choose distinctive names to signal affinity with the wider Black community, the opposite pattern might be expected; Fryer & Levitt, 2004).

The question is: why? As we noted earlier, when surnames are fixed and a population expands, the residual entropy of names invariably rises. In such circumstances, parents may be more likely select diverse first names for their children to increase their name's overall uniqueness. Given the legacy of slavery (Dunaway, 2003), residual entropy is a particular problem for the Black community, where a smaller (less diverse) pool of surnames places more of the burden of providing uniquely identifying information

on first name elements. The residual entropy of Black surnames is considerably higher than for White surnames: the most frequent 500 US surnames convey just 2.5 bits of information about the White community, but over 4 bits of information about the Black community ($t(499)=8.00$, $p<0.001$; Figure 7), meaning that surnames convey far less information about individuals within Black communities. Residual perplexity—which increases when more people share a surname—is twice as high for these names in Black communities than in White communities. Since people called Smith will be likely to give their children distinctive first names (because the surname Smith has high residual entropy), and since this likelihood will increase if a high proportion of their neighbors are also called Smith, the tendency of Black parents to choose more distinctive names for their children when their neighbors are Black than when they are White is not really puzzling at all: it simply reflects the desire of parents to provide each child with a unique identifier.

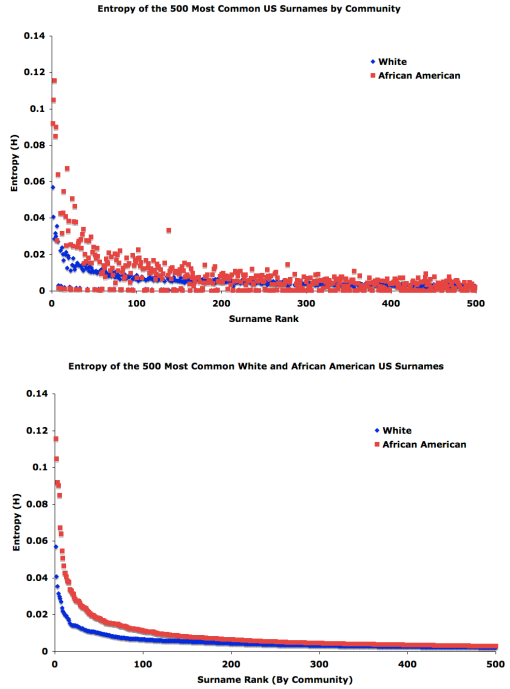


Figure 7. Top: The residual entropy of the 500 most common American surnames in the White- and African-American communities (2000 US Census). **Bottom:** The 500 most common White surnames and 500 most common Black surnames (high residual entropy = less information about individuals).

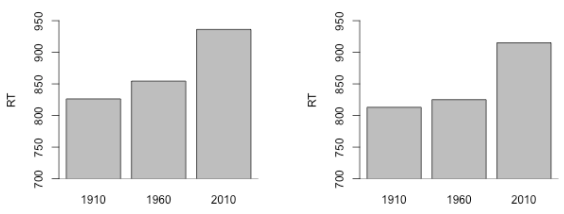
Names, age and memory

Problems with remembering names represent the most disturbing aspect of aging for many people. The analyses reported here raise a question: does memory for names really decline, or are older adults confusing social change with personal change? After all, the changes to the distribution of Western names guarantees that name processing must have grown increasingly difficult over the last century. To provide an estimate of the effect this could have had on name memory, we examined the effect of changing name distributions in a model that simulates human performance in lexical decision tasks.

The naive discriminative reading model (NDR; Baayen, Milin, Durdevic, Hendrix & Marelli, 2011) is a two-layer network that takes letter unigrams and bigrams as cues, and learns to discriminate lexical targets as outcomes (e.g., 'hand', 'John') using the equilibrium equations (Danks, 2003) of the Rescorla-Wagner (Rescorla & Wagner, 1972) learning rule. The model's output is entirely determined by its training corpus—it has no free parameters—and it captures a wide variety of empirical effects in reading (Baayen, 2010; Baayen, Hendrix & Ramscar, in press), and successfully predicts patterns of age-related reading time differences (Ramscar, Hendrix & Baayen, 2012).

To simulate the cross-generational effects of changing name distributions, three versions of the NDR were trained on an identical set of naturalistic linguistic training data (1,500,000 tokens from the Google Unigrams Corpus were used to simulate the experience of reading to age 20). Names from the distribution of US social security applications for a given year (1910, 1960 and 2010) were interpolated into this sample, based on the frequency with which names appear in the corpus, and the distribution givens name in each year. Recognition latencies were calculated for the set of names common to the 1910, 1960 and 2010 name distributions, and for the total set of names learned by each model: Figure 8A shows the cost the distribution of names imposes at each point in time, and Figure 8B shows the average effect this had on precisely the same set of names. The simulations suggest that the simple task of recognizing a name grew harder in the 20th Century, especially in its latter half: the change in simulated reaction time from 1960 to 2010 is three times greater than 1910 to 1960.

Not only did the number of names increase dramatically (the 1960 model learned 60% more names than the 1910 model, and the 2010 model 83% more), but the number of non-name words that were learned declined over time, by 2.5% in 1960, and 5% in 2010. Given that the models were trained on exactly the same number of name tokens, this reflects the degree to which the boundary between English names and non-names has become blurred over time, increasing the memory problem that names pose.

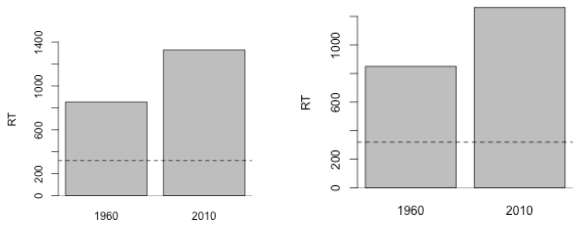


A **B**

Figure 8. A plots the average simulated recognition times for the set of names learned by a "20 year old" model trained on 1.5 million unigrams when sampling from the name distribution in 1910 (left bars), 1960 (center bars) and 2010 (right bars). **B** plots the average simulated reading times in each model for the set of names that are common to the 1910, 1960 and 2010 US social

security applications: i.e., each model's predictions for the same set of names.

Finally, to simulate the effect of these changes within a single lifespan, we compared the predictions of a "20 year old" reading model, trained on 1,500,000 unigram tokens, with names sampled from the summed distribution of social security applications from 1950-1960 (the age at which current septuagenarians were 20), to a model trained on 9,000,000 unigram tokens, sampling from the summed name distribution from 1950-2010 (extending the "experience" of the younger model to age 70).



A **B**

Figure 9. **A** plots the average simulated recognition times for the set of names learned by a "20 year old" model (trained on 1.5 million unigrams, including names from the 1950-1960 distributions; left bars), and a "70 year old" model (trained on 9 million unigrams, including names from the 1950-2010 distributions; right bars). **B** plots the average simulated reading times in each model for the set of names that are common to 1960 and 2010 US social security applications. The area below the dashed line represents a 320 ms response constant (button pressing) added to both models in simulating reaction times.

Names and other proper nouns comprise a very large proportion of natural language: whereas the younger model learned 34,480 word types and 4,540 names, the older model learned 61,839 word types and 19,976 names. Although total vocabulary doubled, name lexicon grew fourfold. Figure 9A shows the predicted impact of experience on average name recognition for someone aged 70 in 2010 as compared to fifty years earlier. Figure 9B shows the projected effect of these processing costs on the same set of names in the same individual. After a response constant is removed from the simulated latencies, the model predicts that on average, simply recognizing a name will take today's septuagenarian nearly half a second longer than when she was 20. Although older adults have hard time remembering names in comparison to their younger selves, a large part of this difference is likely due to the increasing complexity of social name distributions, and the increasing number of names that individuals encounter over the course of their lives.

The Name Game

We identified a common information structure in the world's name grammars that helps satisfy the complex demands of communicating about individuals, and described some of the consequences of recent changes in Western naming practices. Two things are worth noting about these findings: First, the data we report are not inferred from samples of populations, but are instead calculated from records representing the actual populations themselves; and second, the information theoretic methods we used to analyze this data describe and govern all of the physical devices that have come to define our information age. Accordingly, our finding that American female names have twice the perplexity of male names is a statistical fact about the population, which entails that female names in this country must exert considerably higher information processing costs than male names.

These findings may help shed light on many social issues that are far less clear-cut: for example, an often cited reason for the under-representation of women in many professional bodies is that when appointments are made, women's names often don't "come up" (Donald, 2012). It is highly likely that the different processing costs associated with male and female names contribute to this. In a similar vein, these findings offer food for thought for parents choosing names for children in the West, as well as for people with names formed using different grammars as they traverse our increasingly multicultural world. In particular, these findings suggest that the tendency to simply reverse the order of Asian names in Western languages should be viewed with caution.

Finally, these results have implications for our understanding of memory and aging. The belief that memory processes decline as we age rests, in large part, on apparent selective deficits for names in the elderly (Shafto et al., 2007). However, the problem of remembering a name has been getting exponentially harder since before anyone now alive was born (Figure 7). Because current measures of name memory 'deficits' fail to take into account changes to name distributions, it is unclear whether name memory really does decline, or whether these measures simply reflect the overwhelming increases in name information we have documented (see also Dahlgren, 1998; Juncos-Rabadán, Facal, Soledad Rodríguez & Pereiro, 2010). It may be that older adults troubled by their "declining" name memories are presently falling into the trap of taking personal responsibility for a broader social problem.

At the height of the information age, in a world where population growth is inexorably increasing the amount of name information societies must shoulder, the social practices that evolved to maximize the efficiency of name processing are, in many cases, in a state of collapse. Although names are the hardest vocabulary items people have to learn and remember, the problems they pose are currently far harder than they should be. Understanding the information structure of names, as well as how name efficiency can vary and be quantified, can help individuals and societies make more informed choices about names and naming practices. It may even make the names of future generations easier to recall.

For a complete list of **references** and **supplementary materials**, please consult the article copy hosted at *www.sfs.uni-tuebingen.de/*