Sonderdruck aus

Beiträge zur Namenforschung

Band 45 · Heft 2 · 2010

Neue Folge

Universitätsverlag
WINTER
Heidelberg
2010

Inhalt

BEITRÄGE

Harald Jockusch - Alexander Fuhrmann

"Selten gewinnt"- Rare wins. Changes of the frequencies of family names as a consequence of rational choice

Summary: Using German family names as an example, we have simulated the effect of the choice of family names following the rule that the rarer name of the mates is adopted by the children. In Germany there are about one million family names for 82 million people (represented by 28.4 million telephone listings at the reference time, 2005). In this source, about half of the family names occur only once (the "tail" of the distribution, rank R not defined). Assuming asynchronous random mating, and a replacement of each couple by, on the average, two children every 30 years, we find as consequences of the "rare wins" rule:
1. The frequencies F of the most abundant names (R $1 \leq R \leq 20$) decline exponentially with half-lives $t_{1/2}$ of 22 years, regardless whether the "tail" is taken into account or not. Half-lives $t_{1/2}$ increase to >60 years with names of intermediate F (R around 10,000), again with nearly no dependence on the "tail". Very rare names (R around 100,000) with frequencies F of about 2/3 F_{fin} approach the equilibrium value F_{fin} with an overshoot.
2. Without "tail", the most common name (Müller, R = 1, F = 0.96%) approaches F_{fin} after 260 years. With increasing R, times t_{fin} to reach F_{fin} decline steadily to about 100 years for R = 10,000. The "tail" leads to a delay of 22 to 16 years and is therefore of minor influence.

The "rare wins" rule provides a natural and effective means to reduce the frequencies of the most abundant names. It may have applications in countries like Denmark and China in which the ten most common names are borne by 41% and 32% of the population, respectively.

Zusammenfassung: Deutsche Familiennamen wurden als Beispiel verwendet, um den Effekt einer Namenswahl zu simulieren, bei der der seltenere Name eines Paars als Name für die Kinder gewählt wird ("Selten gewinnt"-Regel).
In Deutschland gibt es etwa eine Million Familiennamen für 82 Millionen Einwohner, die im Referenzjahr 2005 durch 28,4 Milionen Telephon-Festnetzanschlüsse repräsentiert wurden. In dieser Quelle sind etwa die Hälfte der Namen nur einmal vertreten ("singuläre Namen", die den "Schwanz" der Häufigkeitsverteilung bilden, in der ein Rang R nicht definiert ist).
Unter der Annahme zufälliger Partnerwahl und des Ersetzens jeden Paares alle 30 Jahre durch im Mittel zwei Kinder ergibt sich als Folge der "Selten gewinnt"-Regel:

1. Die Häufigkeiten F der zwanzig häufigsten Namen ($1 \leq R \leq 20$) fallen exponentiell mit Halbwertszeiten $t_{1/2}$ von 22 Jahren ab, unabhängig davon, ob die singulären Namen berücksichtigt werden oder nicht. Die Halbwertszeiten $t_{1/2}$ nehmen auf >60 Jahre für Namen mittlerer Häufigkeit (R um die 10.000) zu, fast unabhängig von der Berücksichtigung der singulären Namen. Die Häufigkeiten sehr seltener Namen (R um die 100.000, F um 2/3 F_{fin}) überschießen bei der Annäherung an den Gleichgewichtswert (Endwert) F_{fin}.
2. Ohne Berücksichtigung der singulären Namen nähert sich der häufigste Name (Müller, R = 1, F = 0,96%) dem Endwert F_{fin} nach 260 Jahren. Mit wachsendem R verringern sich die Zeiten t_{fin} bis zum Erreichen von F_{fin} kontinuierlich, z. B. auf etwa 100 Jahre für R = 10.000. Der "Schwanz" verlängert diese Zeiten um 22 bis 16 Jahre und ist daher nur von geringem Einfluß.

Die "Selten gewinnt"-Regel ist eine natürliche und effektive Option, um die Häufigkeiten der am meisten verbreiteten Namen zu verringern. Sie könnte in Ländern wie Dänemark und China angewendet werden, in denen 32% bzw. 41% der Bevölkerung Träger der zehn häufigsten Namen sind.

I. Introduction

In China, family names have a tradition of many hundred years; yet, the number of family names is small in comparison to the population size. Thus, each of the three most common family names is borne by 80 to 90 million people, i.e. a population equaling the total population of Germany.

More recently, in most European and European derived cultures, every person has a family name (also called "last name" or "surname"; hitherto abbreviated "name"). Traditionally, upon marriage, the wife adopts the family name of her husband. Consequently, the children propagate their father's name until daughters marry and adopt their husband's name. In this system of "male dominance" the family name is perpetuated independently of its information content; the same would, of course, be true for a system of female dominance. The degrees of freedom for choosing a family name are limited by national name laws (Lerche Nielsen, M., 2007). In recent decades, laws have been changed in most European countries such that the wife's name may become the family and children's name (in Germany Bürgerliches Gesetzbuch §1355, §1616). This freedom of choice enables a rational approach to choosing a family's name. Since the less frequent name has the higher information content, a "Selten gewinnt" ("rare wins") rule appears rational.

The origin and the present frequency distributions of family names in European countries differ according to their histories. The most frequent family names in Germany designate professions (*Müller* - miller, *Schmidt* – smith, *Schneider* – tailor, *Fischer* – fisherman, *Meyer* – administrator, *Weber* – weaver etc. (cf. Kunze, 2004), whereas the most common names in the neighboring Scandinavian countries originally designated family relationships (*Jensen*, *Nielsen*, *Hansen*, *Pedersen*, *Andersen*, *Christensen* etc., with … *sen* meaning “son of …”). While the most common family names in Germany have frequencies around 1%, those of the five most common names in Denmark range from 3 to 5%. Hence, in Denmark, the disadvantage of bearing a very common name, both for an individual and for administrative purposes, has been realized in recent years, and changing a family name into a less frequent, even invented (“face-lifted”; Skadhede, 2007) one has become possible by law in 2006 (Lerche Nielsen, 2007). Although a name frequency of about 1% does not seem very high, practical problems arise e. g. when one searches for a given person *Müller* or *Schmidt* in a telephone directory (e. g. City of Berlin, 3.4 million inhabitants, about 400 names per page, 15 pages for *Müller*, 13 pages for *Schmidt*); in such cases the additional information of first names or at least initials is absolutely necessary but often not sufficient (there are, e.g., 128 *Peter Müller* among the nearly 7,000 *Müller* in Berlin).

Here we suggest a rational choice of spouses’ names upon marriage that avoids the somewhat artificial “face lifting”, i.e. the adoption of invented names, officially proposed in Denmark (Skadhede, 2007). As we show by simulations, adopting the less frequent name of the two spouses leads to a rapid decline of the most common names. A more general treatment of the game “trying to be rare“ has been presented previously (Baake et al., 2006).

II. Material and Methods

2.1 Origin of name lists

The frequencies of family names in Germany were based on the listing of fixed subscribers of the German telephone network (Telecom) in the year 2005.

2.2 Estimating the distribution of German family names

We reconstruct the distribution of German family names on the basis of the following information.

a) the frequencies of the 20 most common names
b) the overall number of names
c) the number of names that only appear from 1 to 17 times
d) 40 frequencies of names distributed in between this unknown range. The best fitting function for this range turned out to be the Hill equation (compare Figure 3),

$$base + \frac{\max - base}{1 + \left[\frac{x_{1/2}}{x}\right]^{rate}}$$

with base = 95641, max = -12.594. rate = 0.8963, $X_{1/2}$ = 16.866.

2.3 Algorithm used for simulation of frequency changes

The algorithm, written in Igor Pro (Wavemetrics), for “marriage” and “propagation” was based on the following assumptions:

The total size of the population is constant. Two individuals collide in intervals of 30 years (i.e. generation time; in Germany this is presently the approximate average age at which individuals marry for the first time). The fertility rate in Germany is about 1.4 children per woman on the average. Also, migration and longer life expectancies alter the population in the future. Since a calculation with these unknown parameters would be very difficult and would change the total number of names, making it complex to analyze the influence of the rational choice of name, we assume a constant population. Our simplifying assumptions enable us to calculate only with the names and their corresponding frequencies. Furthermore, this procedure enables the calculation in step sizes of one year and without distinction according to gender.

The following steps were performed in the simulation:

1. Each individual (as taken from the German telephone listing) is assigned to a group of individuals with identical family names,and the groups are ordered (ranked) according to their size (i.e. the frequency of the family name; cf. 2.2).
2. In a second listing, the individuals are ordered in a randomized fashion.
3. Each 60th individual of the randomized list (from step1) is selected and a second individual (the „mate“) is selected at random. This way, one individual meets another individual, on the average, every 30 years.
4. It is determined which of the two individuals (the “couple“) belongs to the smaller group, i.e. has the less frequent family name.
5. If the first individual belongs to the smaller group, the second individual will change its group (to the smaller one, i.e. changes the name to the less frequent name). Accordingly, if the group of the first individual is larger then this individual will change the group to the one of the second individual. In the rare case that both partners are from the same group (having the same name), nothing changes. In the rare case that both groups are of equal size, the second individual joins the first group. Interactions between members of small groups may lead to the extinction of one group, i.e. a last name, especially when the “tail” of singular last names is included in the simulation.

This algorithm (Steps 2 to 4) is repeated until the end of the list is reached. Then it will start again for the next year of simulation (taking 2 weeks of calculation time on a standard PC for 3,000 years of simulation).

III. Results

3.1 Three cases of frequency distributions of family names: China, Denmark, and Germany

In the following graphs, name frequencies F are plotted against rank R, where R designates the position of a given name when the names are ordered by decreasing F. Relative frequencies F_{rel} are given in percentages of the whole population to facilitate comparisons between countries.

Fig. 1 shows a comparative linear plot $F_{rel} = F_{rel}(R)$ for the twenty most common family names ($1 \leq R \leq 20$) in China (CN), Denmark (DK), and Germany (DE). China is an extreme case, with about 700 family names serving 1.32 billion people. The frequencies of the most common names *Wang*, *Li*, and *Zhang* range between 6 and 7%. In fact, *Wang* is the most common name in the world. Nearly 30% of China's population bears the five most common names, and more than 40%, the ten most common names; the average frequency of all names is 0.14%. In Denmark, with a population of 5.4 million, 21% of the population bear one of the five most common family names - *Jensen*, *Nielsen*, *Hansen*, *Pedersen*, and *Andersen*. And nearly a third of the population bears one of the ten most common names. In comparison to these two countries, Germany's name distribution is moderate. The ten most common names range in their frequencies from nearly 1% (roughly equivalent to 820,000 inhabitants with 260,000 telephone connections) to 0.26 % (Kunze, 2003). The average frequency of names in Germany is about 1/1,000,000.

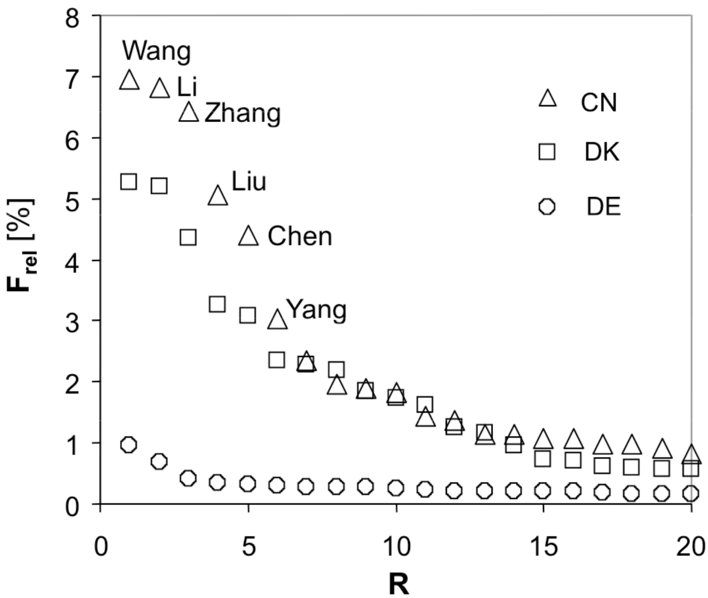

Fig.1: Comparison of the relative frequencies F_{rel} of the most common family names in China (CN), Denmark (DK), and Germany (DE). F_{rel} in % plotted linearly against rank R, with $1 \leq R \leq 20$.

3.2 Basis of calculations and frequency distribution of family names in Germany

The following data and calculations with German family names are based on the Telecom data of telephone connections of the year 2005.

In Germany, with its 82 million inhabitants, the number of family names is surprisingly high. There exists one million family names where, on the average, each family name is represented by only about 80 persons. There were at the reference time 28.4 million telephone connections, so that one telephone connection represents 2.9 people.

Fig. 2 shows the name distribution in Germany for $1 \leq R \leq 20$, as in Fig. 1 but with examples of names given. The 14 most common names are all of German origin and designate professions. Subsequent names designate either professions or personal characteristics like *Klein* (short) or *Schwarz* (black, probably referring to black hair). Although one would expect only one or two blacksmiths per village, and one miller serving several villages, *Müller* and *Schmidt* (the latter representing only the most common of several different spellings) far outnumber the name *Bauer* (R = 14, "farmer").

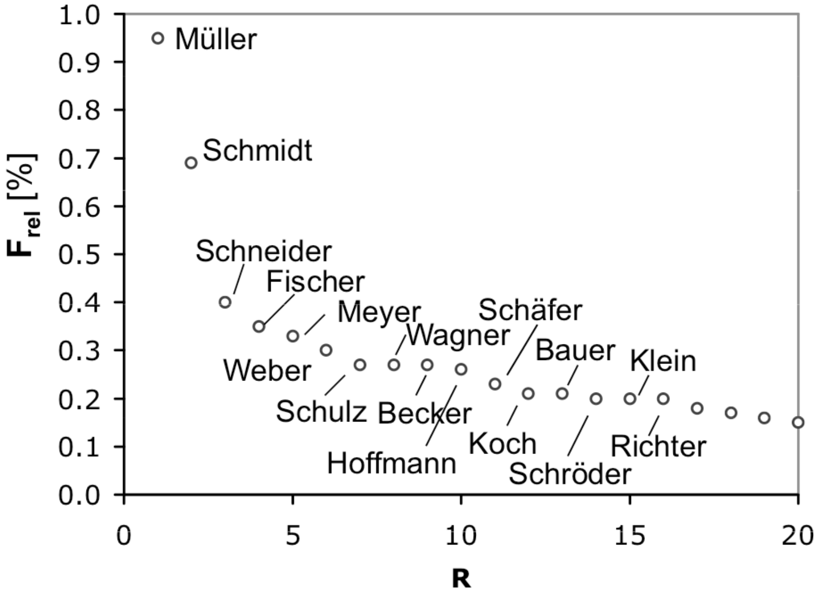

Fig.2: The relative frequencies F_{rel} of the most common German family names ordered by rank R. All shown names except "Neumann" and "Schwarz" designate professions.

The whole range of name frequencies is shown in **Fig. 3** using logarithmically scaled coordinates, again with examples of family names given. Polish (*Kowallik*) and Turkish (*Erdogan*) names show up with intermediate frequencies reflecting the immigration waves in the 19th and 20th century, respectively. The overall log-log frequency distribution is clearly convex and thus does not follow Zipf's law (Zipf, 1935) or any other power law which would have resulted in a linear dependence between log F and log R.

Half of all family names, "singular names", (appr. 500,000) occur only once in the telephone listing. They form the "tail" of the frequency distribution representing 1.8% of the population. Non-German names introduced by recent immigration and individually created double names combined from family names of regular frequencies represent the majority of singular names. With very few exceptions, double names cannot be transferred to the next generation (in Germany: BGB §1617). For this reason, the following simulations were performed both "with tail" and "without tail". The actual case is expected to lie in between these two cases.

3.3 Frequency changes in time as a consequence of the "rare wins" rule

Simulations according to 2.3 were performed to study the changes of name frequencies under the assumption that the "rare wins" rule is followed without exception. Calculations are based on the name distribution in Germany as documented by telephone entries (**Fig. 3**). For our calculations we made the simplifying assumptions that (a) there is no bias with respect to the representation of family names by registered telephone connections and (b) mating is random, "panmictic", and neutral with respect to family names involved. (This is not true for a negligibly small number of partnerships involving nobility).

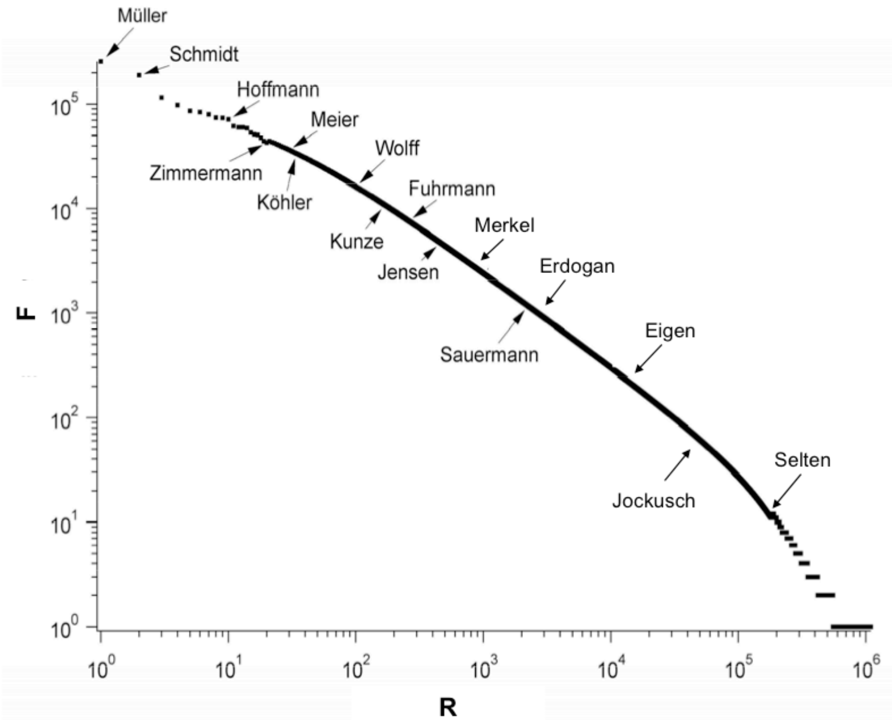

Fig.3: Frequencies F of German family names of all ranks, with singular names (for which R is not defined) included. Scales of F und R logarithmic. Examples of names with widely differing frequencies are given. Jensen is the most common name in Denmark; there is a Danish minority in northernmost Germany. Erdogan is a Turkish name; about 2% of the inhabitants of Germany are of Turkish origin. Kowallik („smith“) is the polish equivalent of Schmidt; there was an immigration from Poland into the Ruhr region in the 19th century. “Selten” means rare. Absolute frequencies F are numbers of entries in the Telecom telephone listing of 2005. The absolute frequency F of *Müller* (about 270,000 entries) corresponds to a relative frequency F_{rel} of 9.5 $^0/_{00}$ (cf. Fig. 2).

In **Fig. 4** the decline of frequencies is shown for R = 1, 2, 3, and 10,000, i.e. the three most common names (Müller, Schmidt, Schneider) and a name with low frequency, in a linear F vs. time plot, with frequencies normalized to initial frequencies (F/F_0). High frequeny names follow an exponential decline, i.e. they are repeatedly and efficiently “diluted“ by the excess of rarer names. The kinetics of a much less frequent name (rank 10,000, frequency 500) is more complex. The decline is delayed by about 60 years (“two generations“) and reaches equilibrium (i.e. average frequency of 43 “without tail“) after about 150 years. For the decline of names with moderate ranks, e.g. up to R = 10,000, half-lives can be defined as the time $t_{1/2}$ at which F(t) is ½ of F(t = 0). Half-lives for names with $1 \leq R \leq 10{,}000$ are shown in the **inset** of **Fig. 4**. They amount to 22 years independently of R for low Rs and increase to 35 years for R = 1,000 and 66 years for R = 10,000 with practically no dependence on the „tail“ in this range of Rs.

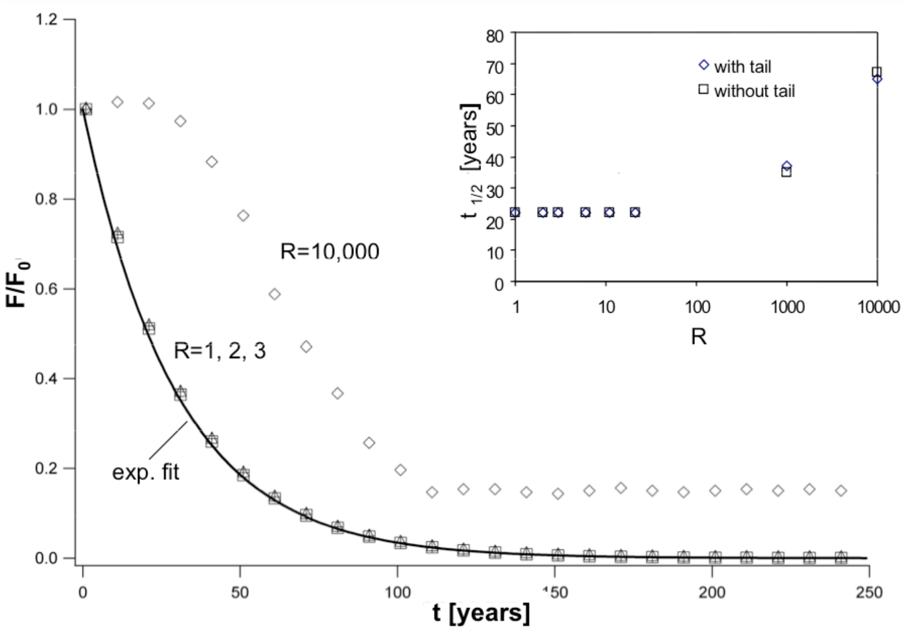

Fig.4: Decline of the frequencies of the three most common family names R = 1, 2, 3 (Müller, Schmidt, Schneider, cf. Fig. 2) and a name of intermediate frequency (R = 10,000) as a consequence of following the rule “rare wins”. Simulation is without „tail“. For low R, the decline is exponential. For intermediate F (example F = 10,000), the decline is preceded by a lag.
Inset: Half lives of name frequencies as a consequence of the “rare wins” rule. Half lives $t_{1/2}$ plotted against ranks R (log scale). Simulations are without and with “tail”. For low R, $t_{1/2}$ are about 22 years and independent of R. For high Rs, half lives increase to about 70 years and for very high Rs half lives are not defined. There is only a negligible influence of the “tail”.

Very rare names, apart from the singular names, are those for which the frequencies are close to or below the average frequency of 43 telephone registrations ("without tail", roughly corresponding to 125 persons or 0.0015 $^0/_{00}$ of the population). Examples of kinetics F vs. time with selected starting values of F_0 from 2 to 2,400 are shown in **Fig. 5**. Though names with frequencies below 20 monotonically reach the equilibrium value of 43 after 50 to 120 years, those with initial frequencies around 25 to 30 show an overshoot.

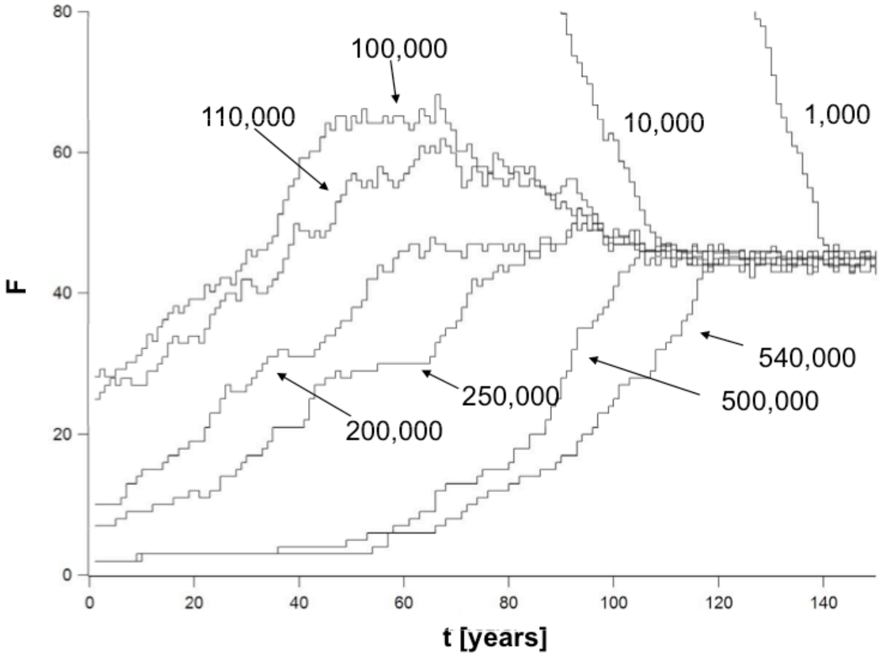

Fig.5: Approach of the equilibrium value F_{fin} by names of intermediate (1,000 ≤ R ≤ 10,000) and low frequencies (100,000 ≤ R ≤ 540,000, starting F < 30). Some of the latter show an overshooting approach to the equilibrium value F_{fin}. Simulation is without tail.

An overview over the whole frequency range is shown in **Fig. 6** using a logarithmic scale for the frequencies.

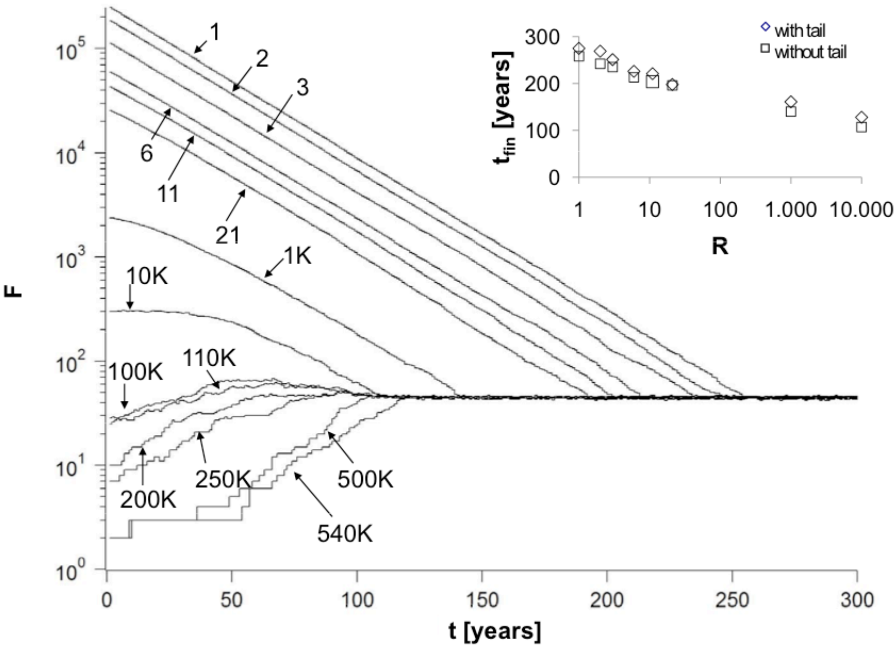

Fig.6: Approaching the equilibrium F_{fin} by names covering a wide range of ranks R. Frequencies F (left) and F_{rel} (right) plotted logarithmically. Linearity for $1 \leq R \leq 21$ indicates exponential decline. The most common name, Müller, takes about 255 years to reach the equlibrium value. Simulation is without tail. K stands for thousand.
Inset: Times t_{fin} to reach the equlibrium F_{fin} as a function of rank R. Simulations are with and without “tail”. The slopes of the curves are similar, but the presence of the “tail” delays the approach to the equilibrium by 16 to 22 years (depending on R) due to the fact that the final value (listings per name) is about half in the presence of the “tail”.

Names with high frequencies reach equilibrium after a t_{fin} of 200 to 260 years, whereas the lowest frequencies reach equlibrium after 60 to 80 years (**Fig.6 inset**). The value of t_{fin} depends on whether the „tail“ is taken into account or not. While the steepness of the decline of t_{fin} with increasing R is similar for both cases, t_{fin}s are 16 to 22 years longer when the “tail” is taken into account. This difference is less than a generation time and therefore not relevant for the overall result of our simulation.

IV. Conclusions

We have characterized the distribution of the frequencies of family names in Germany based on the telephone listing of 2005. Due to the increasing use of cellular phones and unlisted telephone connections, it will become more and more difficult to use telephone listings as a representative source of information in the future.

The frequency distribution of the one million family names of the inhabitants of Germany is characterized by a convex log-log shape followed by a “tail“ of about 500,000 singular names. The bearers of these names are probably very heterogenous and their contribution to the future composition of the German population is difficult to evaluate. We have therefore performed our simulations in parallel with and without „tail“. It turned out that in most aspects the results are only marginally different, especially for common names and the initial kinetics of their frequencies. This seems plausible as the bearers of singular names constitute only 1.8 % of the population (500,000 for 28 million telephone listings) and thus the probabilty to get assigned to one of these as a partner is correspondingly low.

It should be pointed out, however, that our kinetic simulation is based on a rather simplified model of the population and its reproduction. We have assumed a reproduction rate of two children per woman and an immediate replacement of “parents” by their “children” resulting in a constant population size. In fact, the present reproduction rate in Germany is only about 1.4 children per woman, people get progressively older, and the population size is decreasing. The rates of frequency changes resulting from our simulation are consequently overestimates and the half lives of 22 years of the most common family names are underestimated.

We wish to make the point, however, that due to the initial steepness of the exponential decline, it is realistic for a person with a very common name to experience the drop in frequency of that name and gain informational value within a lifetime.

The problem of very common names in Germany is moderate as compared to the situation in Denmark or China. In Denmark there is public concern about extremely common names and legal action has been taken to allow the invention of family names (“face lifting of family names“; Skadhede, 2007; Lerche Nielsen, 2007). We propose a more “natural“, gentle method based on the legal possibility that a family may assume the husband’s *or* the wife’s name. If, upon marriage, a couple would always assume the rarer name, the frequencies of the most common names would drop noticeably within a lifetime.

It would be worthwhile to apply our model and simulations to countries like Denmark and China in which extremely frequent family names occur.

Dedication and Acknowlegements

This paper is dedicated to the memories of Heinz Sauermann (1905-1981), professor of economics at Frankfurt University and fatherly friend of H. J., and of Karl Peter Grotemeyer (1927-2007), professor on mathematics and rector of Bielefeld University for 23 years.

Work has been presented at the Manfred Eigen Winter School at Klosters, January 2009, and the CAS-MPG Partner Institute for Computational Biology, Shanghai, August, 2009. The authors thank Professors Konrad Kunze (Freiburg, Germany), Ernst-Martin Füchtbauer (Aarhus, Denmark), and Michael Lerche Nielsen (Copenhagen), and Carl G. Jockusch (Urbana, Ill.) as well as Drs. Quiang Li (MPI for Demographic Research, Rostock Germany), Wolf Jockusch (Carl Zeiss Microimagung GmbH, Göttingen), and journalist J∅rgen Skadhede (Jyllands Posten) for valuable informations, suggestions and help, and Professor Reinhard Selten for not objecting to the use of his family name (*nomen est omen*).

V. References

M. Baake – U. Grimm – H. Jockusch (2006), Freely forming groups: trying to be rare The ANZIAM Journal 48, p. 1-10

BGB = Bürgerliches Gesetzbuch

K. Kunze (2003), Namenkunde. Vor- und Familiennamen im deutschen Sprachgebiet. 4th edition. dtv, München, 2003

M. Lerche Nielsen (2007) Et ar med den nye danske navnelov. Stud. Anthroponymica Scandinavica 25, p. 95-117

J. Skadhede, Danskerne giver deres navne plastikkirurgi. Jyllands Posten 19. 5. 2007

G. K. Zipf (1935), The Psychobiology of Language. Houghton-Mifflin

Adressen der Autoren: **Prof. Dr. Harald Jockusch**
h.jockusch@uni-bielefeld.de
Tel.: ++49 7664 408004
Oberes Breitle 3
79112 Freiburg
Developmental Biology and Molecular Pathology, Bielefeld University, D-33501 Bielefeld, Germany

Alexander Fuhrmann, PhD
Alexander.Fuhrmann@asu.edu
Department of Physics, Arizona State University, Tempe, USA