Are new words predictable?
A pilot study on the origin of neologies by means of natural selection

Dietmar Zaefferer
Ludwig-Maximilians University Munich

20 November 2019

Chapter to appear in Fiorentini, I., Goria, E., & Mauri, C. (2020). *Building Categories in Interaction: Linguistic Resources at Work*. New York: Amsterdam: John Benjamins.

Address correspondence to: Dietmar Zaefferer
Theoretical Linguistics
Ludwig-Maximilians-Universität München
Schellingstr. 7 D-80799 Muenchen
zaefferer@lmu.de Germany
Phone: +49 89 366675

Abstract

Facing an unknown phenomenon agents have to create a new concept, and if they want to communicate about it they have to find a fitting label. They do so by recombining or figuratively stretching familiar concepts and by creating new expressions or by rearranging or enriching already available elements. In general this kind of process is studied in hindsight, looking at neologisms that have emerged in the past. But linguistic theorizing is about explanation and explanation is tested by prediction, so this study probes into the predictability of this phenomenon using an experimental approach. Elaborating on ideas presented by Müller (1870), approvingly taken up by Darwin (1871) and specified among others by Croft (2000) and Aronoff (2017) the theoretical basis of the investigation views the dynamics of thought and language as special kind of evolutionary process. Given the hybrid nature of language as a quasi-natural social artifact, the creation of a novel concept and the spread of a new label or usage for it can be assumed to combine conscious reflection and more or less unconscious processes of adaptation both within and across individuals. The research question to be answered by this study was: What is the identity and relative weight of the factors that determine the conceptualization of a new phenomenon, the creation of different labels for it, its ranking by individuals and by a community and eventually the selection of the survivor? The results are surprising in two seemingly incompatible regards: One the one hand subjects produced a stunning pool of different labels and were thus far more creative than anticipated, one the other hand the derived properties of the most successful competitors were sufficient to demonstrate that approximately predicting the survivors is nevertheless possible.

Contents

1. Introduction

1.1. Predictive language theories and their moving target

Arguably, one of the most fundamental questions in linguistic theorizing is this: 'Why are human languages the way they are?'. The correct answer is of course easy to find: 'Because they became to be that way.' But this answer only leads to the next question: 'How did this happen? What are the forces that determined their prehistory and history up to their present stages?' And this is a deep question that possibly will never find a complete answer. But scientific progress consists in giving provisional and partial answers, answers that slowly are moving towards higher degrees of reliability and completeness.

Whereas the evolution of full-fledged human languages from more primitive communication systems is still bound to be highly speculative (in spite of the considerable progress that has been made in the last decades[1]), the constantly ongoing dynamics of languages is open to less conjectural scientific investigation: Based on the observation of changes in the past theories can be devised in order not only to explain those historic changes and similar changes in progress, but also to predict changes to come.

Research on neologisms has so far been dominated by retrospective investigations (cf. e.g. Schmid 2008). With the expansion of the World Wide Web the detection of new words has become an interesting strand of neologism research (cf. Kerremans and Prokić 2018). In both cases only pre-existing products of language competence are being investigated. Only very recently the concept of 'Predictive Lexicology' (cf. Veale and Butnariu 2006) has entered the stage in the domain of computational creativity. By contrast, the study to be presented below is to do with natural creativity, and as such it is probably the first to directly tap into the linguistic competence by eliciting new names for unfamiliar entities.

Similarities with biological evolution have already been pointed out by Darwin, but the helpfulness of this analogy is still a controversial issue (cf. Pinker and Bloom 1990, Mesoudi et al. 2006, Progovac 2019). The present pilot study proves that at least in the area of neologisms the comparison with biological natural selection fosters theory building in that it leads to falsifiable predictions about the fate of names for new phenomena. It does so by interpreting fitness as preference in the context of selection from a pool: Given a set of competing labels for a new category, the one that survives as a result of its being accepted by the community can be assumed to be the most preferred one. Conversely, starting from a theoretical preference ranking it should be possible to predict acceptance by a community at least to a certain degree.

These considerations naturally lead to the following research question: What are the factors that determine the preference ranking of competing labels in a community and hence the selection of the survivor? Once they are found, new words become predictable in the sense that in a given pool of candidates for naming a new phenomenon, those that are more likely to win the competition can be distinguished from the others.

[1] Cf. the proceedings of the biennial International Conferences on the Evolution of Language (EVOLANG) that took place twelve times so far since 1996.

If the theory is successful it still leaves open other aspects of the predictability of new words, first and foremost the emergence of new phenomena to be labeled (which of course for the most part is a question linguistic theory is neither called for nor capable to answer), and second the factors that determine the creation of the pool (which is a question only a joint effort of linguistics and psychology can answer), but at least a first step will have been made towards a more comprehensive theory of the ways neologisms emerge, and a proof of concept will have been given that making new words predictable is not a completely hopeless enterprise.

1.2. Language change and neology

Languages and words are artifacts of a special kind insofar as in everyday life they are perceived as something natural rather than artificial. On the other hand, words are those elements of language whose nature as artifacts becomes most easily visible, especially when new categories require the creation of new linguistic forms or of new readings of existing ones. This is regularly the case when new phenomena such as new artifacts are to be dealt with in a community and become important enough for the need to agree on its categorization and linguistic coding to arise.

Given this hybrid nature of language as a quasi-natural social artifact, the creation and spread of a label for an unfamiliar category can be assumed to combine conscious reflection and unconscious processes of adaptation both within and across individuals.

The central idea of the investigation to be presented here was to probe experimentally into the identity and relative weight of those factors that determine the form and early fate of words created for communicating about novel categories. It is a pilot study in the sense of a pretest of an unprecedented research procedure.

The dynamics of natural living languages can be compared to that of glaciers: at first glance they look motionless, but upon closer inspection they turn out to be in constant motion, snowfalls add new substance which gets transformed into ice, the ice moves slowly downward, ice blocks fall from their terminus, and substance gets lost as melt water runs off.

Similarly in the body of a living language four regions corresponding to four kinds of language change can be distinguished:

1. The core region comprises those elements that stay the same in the slow movement of the whole: *No language change* is the limiting case of linguistic dynamics, and it is of crucial importance for the functionality of language[2].

2. The lateral regions contain those elements that undergo modification: *Transforming language change* applies to either form or function or both form and function.

[2] "Every language changes at a rate which leaves contemporary persons free to communicate without disturbance." (Bloomfield 1926: 162)

3. The lower region holds the elements that go out of existence ('melt away'): *Subtractive language change* is to do with the forms or functions or form-function pairs that become obsolete and finally disappear.

4. The upper region is where elements come into existence ('new snow accumulates'): *Additive language change* happens when fresh elements enter the language.

The focus of this paper, of course, is on additive change, but the other regions are relevant as well: Most morphemes used in neologisms come from the stable core and move to the lateral regions of transformative language change in that their meaning potential is modified by the addition of a new, figurative usage. Finally, as is well known, successful new words that compete with pre-existing ones can push down the former to the lower region by marginalizing and ultimately ousting them (subtractive change).

An important question that forms the backdrop of all neologism research is: Why do new elements of language emerge at all? I submit that the most general answer is one that is best illustrated by an economical metaphor: because there is demand for them. Although demand can also come from various minor sources such as taboo avoidance,[3] the strongest motivation for demand is provided by new phenomena that come into existence, requiring thus that new categories be built for dealing with these phenomena. In order to communicate about such novel categories, in the beginning deictic expressions ('the thing over there') or lengthy phrases ('the gizmo with a blue ball on top and three feet') will suffice, but in line with Zipf's Law this becomes too cumbersome as soon as communication about a new category occurs sufficiently often and becomes sufficiently relevant for a community. This is when demand for a compact way of dealing with the concept exceeds the relevant threshold value and a new linguistic element is born.

These assumption lead to the first working hypothesis the present study is based on:

> ***H1 Economy Hypothesis***: Neologisms are supplied if and when there is sufficient demand.
>
> The Economy Hypothesis leads directly to our central research: how is this demand satisfied? The answer to this question is given by our second working hypothesis:
>
> ***H2 Selection Hypothesis***: Neologism supply takes place in two steps: (a) a candidate pool of possible names is generated, and (b) a selection process eliminates all candidates with an insufficient degree of fitness such that only one (rarely more) candidate survives and wins.[4]

[3] Cf. e.g.: "It is a likely speculation that the Norman French title Count was abandoned in England in favour of the Gemanic Earl […] precisely because of the uncomfortable phonetic proximity to *cunt*" (Hughes 1991: 20).

[4] This is the well-known avoidance of narrow synonyms ("je ne crois pas qu'il y'ait de mot synonime dans aucune Langue [I do not believe that there is a synonymous word in any language]." Girard 1718: xxx [roman number 30]). In line with our approach there have been attempts at explaining this avoidance in evolutionary terms, e.g., by Aronoff (2016), who draws on Gause's principle of competitive exclusion: "... as a result of competition two similar species scarcely ever occupy similar niches ..." (Gause 1943: 19); cf. also Goldberg 2019: 26.

At this point a new question arises: Is this artificial selection or natural selection? Before addressing this issue it seems appropriate to embark on a short excursion into the history of science regarding the relation between Darwinian evolution and language change.

1.3. Language evolution: from Darwin via Haeckel, Schleicher and Müller back to Darwin

In November 1861 the famous and highly influential German biologist Ernst Haeckel[5] "buries himself" (his own words in a letter to his fiancée) in the German translation of Darwin's (1859) masterpiece *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life* (Richards 2008: 68). Excited about the new ideas he urges his friend August Schleicher, the famous linguist (who earlier had introduced tree diagrams for language families) to read it too. Schleicher, immediately excited as well, reacts with an open letter to Haeckel that comments on the intriguing analogies between the evolution of species and the evolution of languages (*Die Darwinsche Theorie und die Sprachwissenschaft* [Darwin's Theory and Linguistics], Schleicher 1863).

The loop circles back to England when Schleicher's text is translated under the header "Darwinism tested by the science of language" (a title echoed by Aronoff 2017) and published in 1869. Max Müller, the famous German philologist in Oxford, reviewed it approvingly in the very first number of *Nature* (Müller 1870). Most interestingly in view of the present topic, Müller points out a parallelism overlooked by Schleicher: "A much more striking analogy, therefore, than the struggle for life among separate languages, is the struggle for life among words and grammatical forms which is constantly going on in each language. Here the better, the shorter, the easier forms are constantly gaining the upper hand." Moreover, Müller reveals himself as a functionalist: "... though we cannot in every instance explain the causes of victory and defeat, we still perceive, as a general rule, that those words and those forms carry the day which for the time being seem best to answer their purpose." (Müller 1870: 257)

In *The descent of man, and selection in relation to sex*, his second landmark book on evolutionary theory, Darwin includes a section on language (1871: 85ff.), where he quotes (somewhat inaccurately) from Müller's review of Schleicher's open letter:

> "We see variability in every tongue, and new words are continually cropping up; but as there is a limit to the powers of the memory, single words, like whole languages, gradually become extinct. As Max Müller[69] has well remarked: —'A struggle for life is constantly going on amongst the words and grammatical forms in each language. The better, the shorter, the easier forms are constantly gaining the upper hand, and they owe their success to their own inherent virtue.' To these more important causes of the survival of certain words, mere novelty and fashion may be added; for there is in the mind of man a strong love for slight changes in all things. The survival or preservation of certain favoured words in the struggle for existence is natural selection. [[69] 'Nature,' January 6th, 1870, p. 257.]"

This should suffice for motivating the short excursion into the history of science. Almost all the relevant topics of the present paper are already addressed here: the perpetual emergence of

[5] Famous among other things for coining the neologism 'ecology' (Haeckel 1866: 286).

lexical innovations as well as the causes of the survival of certain words, such as the preference for better, shorter, easier words (Müller), and the love for novelty and fashion (Darwin) at work in the lexicon. In honor of these outstanding scholars the former idea will be called 'Müller's Hypothesis' and the latter one 'Darwin's Hypothesis.'

H3 Müller's Hypothesis: In their constant struggle for life the better, the shorter, the easier words and forms are the ones that survive.

H4 Darwin's Hypothesis: In addition to quality, briefness and ease mere novelty and fashion contribute to the survival of words and forms.

One more topic that is crucial for the present discussion is addressed by Müller, who in his Schleicher review adds the following caveat (1870: 258):

> "But these analogies should not be carried too far. At all events we should never forget that, if we speak of languages as natural productions, and of the science of language as one of the natural sciences, what we chiefly wish to say is, that languages are not produced by the free-will of individuals, and that if they are works of art, they are works of what may be called a natural or unconscious art – an art in which the individual, though he is the agent, is not a free agent, but checked and governed from the first breath of speech by the implied co-operation of those to whom his language is addressed, and without whose acceptance language, not being understood, would cease to be language."

This takes us back to the question asked at the end of the preceding section: Is the selection of neologies artificial selection or natural selection? Darwin's answer is unmistakable: "The survival […] of […] words […] is natural selection." But Müller warns us that biology and works of art are different in that only the latter are the product of agents, and that the common denominator in the selection processes is only the absence of consciousness. But is human agency not also the hallmark of what inspired Darwin's theory, namely artificial as the opposed natural selection? Does Müller not imply that what is at work in language is in a way not really natural, but artificial selection?

And indeed, a closer look at Darwin's terminology seems to support this view. Discussing the category of artificial selection in *The origin of species* he makes an often ignored important distinction (emphasized by Sterrett 2002) between what he calls 'methodical selection,' which is based on conscious, intentional action, and what he calls 'unconscious selection,' which although also being based on conscious, intentional action, is an unintended consequence of it on the social level:

> "I can see no more reason to doubt this [that the swiftest and slimmest wolves would have the best chance of surviving], than that man can improve the fleetness of his greyhounds by careful and *methodical selection*, or by that *unconscious selection* which results from each man trying to keep the best dogs without any thought of modifying the breed." (Darwin 1859: 90f.; emphasis added)

Since the creation, selection and survival of new words is clearly not a matter of biological, but of cultural evolution (cf. Richerson and Boyd 2005, Mesoudi et al. 2006), the term pair methodical selection and unconscious selection could be adopted, the former for cases of intentional choice (e.g. "The government doesn't like 'inequalities' in health. Even the word is

banned: 'variations' is the acceptable word." Smith 1997:51), and the latter for processes in which several competing names are used in a community until a winner emerges.

But since there is a common denominator of biological and cultural natural selection in that both are non-intentional uncontrolled processes, I will not speak of methodical selection and unconscious selection, but rather extend the use the term pair artificial selection and natural selection to the domain of cultural evolution, seemingly vindicating Darwin's verdict that the survival of words is natural selection (although he never wrote about cultural evolution). The resulting picture comprises thus four categories: Artificial and natural biological selection on the one hand, and artificial and natural cultural selection on the other. The use of 'natural selection' in the title of this paper means of course the last mentioned category.

1.4. The incredible career of a neologism and Engelbart's Question

During presentations of preliminary results of this study the author used to challenge his audience as a little warm-up exercise with the following question and picture:

What is this?

Figure 1

Since nobody knew the correct answer, he gave it himself, first in the following form:

(1) *x-y* position indicator for a display system.[6]

Then he presented other possible formulations of the correct answer:

(2) position indicator for a display system
(3) visual display input device
(4) cursor position controller for a display device
(5) apparatus for controlling movement of a curser [sic!] in a computer display system
(6) optical cursor control device
(7) cursor control device for use with display systems[7]

None of them seemed to be especially enlightening. So one more answer was added that solved the riddle:

(8) computer mouse

[6] This was the name used by Engelbart in patent US 3541541 A, filed on Jun 21, 1967. (https://patents.google.com/patent/US3541541A/en)
[7] All examples are titles of other patents to be found in the text of the abovementioned patent as citing it.

Together with the device it denotes this neologism made an incredible career: in 1981 the compound *mouse click* (the second member was chosen for disambiguation) appeared only once in the 46,107 books Google's Ngram Viewer is based on for this year, twenty years later 'mouse click' occurs as many as 2,535 times in the 104,147 books sifted through in 2001.

How did this incredible career start off? Here is what is known about it:

In the early sixties Doug Engelbart and his collaborators at Stanford Research Institute were testing a two-wheeled wooden box (figure 1) that was supposed to complement the functions of the computer keyboard and that later was to be patented under the formal name '*x-y* position indicator for a display system.' Several names were proposed.

> " 'We set up our experiments and the mouse won in every category, even though it had never been used before,' said Mr Engelbart. 'It was faster, and with it people made fewer mistakes. Five or six of us were involved in these tests, but no one can remember who started calling it a mouse. I'm surprised the name stuck.' " (Beaumont 2008: 1) " 'We thought that when it had escaped out to the world it would have a more dignified name,' said Mr Engelbart. 'But it didn't.' " (ibidem)

Engelbart had no viable theory of the form and fate of names for new categories. With such a theory he wouldn't have been surprised that the name stuck. In honor of Douglas Engelbart, the underappreciated pioneer of augmenting the human intellect and boosting collective IQ (Engelbart 1995) who died in 2013, the following neology based on his name will be used:

Q1 Engelbart's Question: Why do some neologies stick, whereas others don't?

Using this neology the purpose of the present study can also be characterized as trying to give an empirically supported answer to Engelbart's Question.

1.5. Naming and categorization: clarifications and sample case observations

The story of the computer mouse and its collective christening is told here among other things in order to give a real-life illustration of the issues to do with the relation between categorization and naming, both individually by an agent and interactively by a group. Before coming back to the example a few clarifications may be helpful in preventing possible misunderstanding.

First, it is worth emphasizing that naming and categorization are not the same. As the focus of this paper is on naming it is certainly not central to the topic of the present volume. But since categorization is difficult to access without looking at language, and since naming and categorization are closely related the interactive naming processes studied here will offer also some insights into the categorization processes involved.

Second, whereas categorization is most of the time understood as assignment to a given category it is clear that this volume is rather to do with categorization in the sense of category formation, i.e., the creation of entities that allow for categorization in the other sense to occur in the first place. The same holds for the senses of naming: in most contexts, for instance in psychological naming tests, the task consists in finding a pre-existing name that fits the stimulus. In the present context naming is the creation of new names, i.e., name formation.

Third, concepts and categories are taken here to be abstract entities that are different both from their instantiations (or sets thereof) and their (neural) representations in individual minds. Instead the following definition will be assumed: for an agent or a group of agents to dispose of a concept or category means being able in the context of a given purpose (a) to distinguish between instantiations and non-instantiations of the concept and (b) to neglect the distinctions between different instantiations of it. A very simple example is the thermostat, which disposes of three temperature categories: inside the desired range, above it, and below, because it regards the difference of temperatures in distinct categories and disregards temperature differences within the same category.

Returning to the case of the computer mouse it obviously belongs to the category of artifacts and more specifically to the tools. As with all tools its closer categorization has to be in functional terms. Since the task of Engelbart's team was defined in functional terms (to develop a device that allows for controlling the position of a movable mark on a computer screen) there can be no doubt that every team member disposed of a clear and shared concept of it. The words that were used to refer to it before being ousted by the animal metaphor are not known, what is known, however, is that there was a competition not only among expressions, but also among devices: "A total of five different hardware devices were tested […] The devices included the light pen, and four types of bug-positioning devices: a joystick, a 'mouse', a Grafacon, and a knee control." (English et al. 1965: 41)

Interestingly, the quote mentions another neology, 'bug' in the sense of the mark on the screen the positioning device points at. Since it had the shape of a plus sign, the motivation by similarity with a little animal is quite transparent.[8] In this hardware competition situation for a team member to speak of a bug-positioning device was not enough to refer specifically to what was later to be called a mouse: bug-positioning device is the strongest common superordinate concept (*genus proximum*), but there was demand for a more specific concept with a distinctive feature *(differentia specifica*) to tell the new device developed by the group from its competitors. Given the cubical shape of the first mouse (see figure 1) one can speculate that something like 'the box device' may have been used. Interestingly, the animal metaphor was so successful that already in the abovementioned technical report the nickname "mouse" was used together with the other rather technical designations.

But what could have inspired its creation? In view of the original box shape that was quite different from the round bodies of later mice, the form was probably not the main motivation. The association of the flexible cable coming out from the device's body with the tail of a mouse seems plausible, but since the competitors were connected in the same way the presence of a cable was not distinctive. In the absence of more information every plausible explanation is bound to be speculative, including the following guess: perhaps there was a double motivation by two conspiring factors, the comparatively small size and the motion that it causes in the 'bug.' Given that a mouse is a prototypically small animal and that mice do eat bugs that could have been a motivation. An additional factor may have been the iconic

[8] Nevertheless, as is well known, this use of the insect metaphor was later displaced by the concept of a defect in either hard- or software.

relation between sign and signified constituted by the monosyllabicity and hence minimal size of the word and the tiny size of the animal. This may be the answer to Engelbart's Question in this special case.

And of course after the group consensus about the name had been achieved the spread of its use beyond the Engelbart lab may well have been supported by factors such as the resemblance of the cable and a tail, the similarity of the shape of later mice with that of the animal, and the sympathy many people have for mice (as opposed to rats).[9]

For the present purpose of developing a predictive theory these observations on a sample case may be more or less helpful depending on its degree of prototypicality. In order to get beyond the single case a large amount of additional cases had to be analyzed.

2. Investigating form and fate of neologisms: an experimental approach

2.1. Creating demand for names of unknown objects: meet the Fribbles

In order to come up with an empirical and generalized answer to Engelbart's Question, and in view of the fact that it is impossible to wait in the lab until a new phenomenon appears in the public domain, a new experimental kind of neologism research had to be devised. In the first place, under the assumption that H1, the Economy Hypothesis is correct, the following question had to be answered: how can an experimenter artificially create demand for neologisms in his subjects?

To begin with, stimuli had to be found that were completely unknown and new to the participants. Fortunately there exists a whole family of this kind of stimuli, the artificial animal-like objects called *Fribbles* (created by Michael J. Tarr, Brown University, www.tarrlab.org; cf. also Barry et al. 2014).[10]

Five of the 36 Blue Fribbles (the three most-different exemplars of all 12 species) were selected:

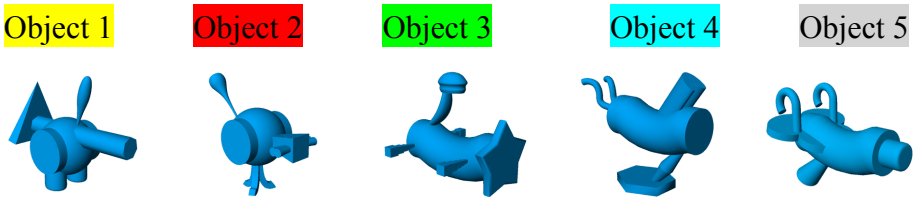


Figure 2

With their help it is easy for the experimenter to create demand: all he has to do is to instruct the subjects to think up names for these objects.

2.2. Preparatory data collection: five objects, five name pools and an initial hypothesis

[9] Other motivations are of course possible. At the Bologna meeting the present volume is based on Larry Barsalou (p.c.) commented that he associates the device with the animal because both typically scurry to and fro.
[10] The author had been made aware of the Fribbles' existence a decade ago by Daniel Casasanto, who had used them successfully himself (Casasanto 2009a).

The project took off with a rather informal first step: nine students (fluent speakers of German, six females and three males) of a seminar on essential statistics for linguists were given twenty minutes to fill a one-page questionnaire with the header 'Naming of unknown objects,' showing the five object numbers as column headers and three empty lines below each one. Slides of the objects were projected for approximately four minutes each. Subjects were asked to come up with three possible names they would give the relevant object, without talking to each other. There was no remuneration. Out of the 135 slots, 55 were filled with mostly novel German words. Although there was considerable overlap concerning the roots (the eleven proposals for object 3 all contained the German root *Stern* 'star'), there was no overlap at all among the full names.

The preparatory pool for object 1 contained 17 names, among them *Hammerschweinchen* 'hammer piglet' and *Sparschweinkanone* 'piggy bank cannon', with six occurrences of the root 'pig' and three occurrences of the root 'hammer.'

Object 2 yielded only 9 responses, including *Kastenkugel* 'box sphere,' *Kopfhörerhund* 'earphone dog' and *Ohrenvogel* 'ear bird.' The root 'bird' occurred three times, the roots 'ear' and 'sphere' twice.'

With eleven different responses the harvest inspired by object 3 was slightly bigger. As already mentioned, all used the root 'star' as component, ten in modifier position (e.g. *Sterngurke* 'star cucumber,' *Sternstempel* 'star stamp'), and one as head: *Aussichtsstern* 'vista star.'

Object 4 elicited nine words, and there was no overlap at all, neither of words nor of roots. Examples are *Schneckenvogel* ('snail bird') and *flügelloser Schmetterling* ('wingless butterfly').

Object 5 was again given nine names, among them *Hakenzahnpastatube* 'hook toothpaste tube' and *Aufhängesitz* 'hang up seat'. The roots 'hang' and 'tube' occurred three times each, the root 'hook' twice.

2.3. Venturing a first hypothesis

The recurring features of the 55 items in the five preparatory pools were encouraging: they hinted at the possibility that new words are predictable at least in their form (noun-noun compounds are preferred) and the general semantic categories involved. Most roots denote either animals (pig, elephant, bird, snail) or artifacts (hammer, hook, tube, seesaw), and shape-denoting roots (star, sphere, hook) seem to have an important role as well.

Therefore the following tentative hypothesis was formulated:

H5 Initial Word Prediction Hypothesis

German names for novel objects are preferred to the degree that they meet the following criteria:

- they have the form of a binary NN compound,
- the head of the compound denotes an animal or artifact and

• the modifier of the compound has a shape-related meaning.

Fitting examples from the five preparatory pools are *Dreiecksstempel* ('triangle stamp'), *Rundtröte* ('round toy trumpet'), *Sternsaurier* ('star dinosaur'), *Schneckenvogel* ('snail bird'), *Hakenhebe* ('hook lifter').

But the preparatory pools were of course by far too small in order to derive predictors for an experiment. In addition to poor size, another shortcoming of the pools was that they originated from the activity of isolated subjects, whereas the creation of new words is not restricted to individuals, rather it is probably most of the time or at least quite often a social process where ideas are exchanged and proposals are compared. And finally, speakers may have a good sense for the difference between their own preferred words and those neologisms that are likely to be accepted in a community.

Therefore a pretest was designed in order to get rid of these shortcomings by means of involving group processes and of tapping into the intuition of speakers about social preferences that are possibly distinct from their own particular likings.[11]

3. Pretest: deriving preference hypotheses from estimated natural selection

3.1. Pretest procedure

Five subjects were recruited with the help of the mail service 'Infodienst' of Ludwig-Maximilians University in Munich, three females and two males. They received 9 Euros each for their participation. The session lasted for about 60 minutes and took place in the university's MELESSA lab; it was recorded with the lab's audio and video devices. The experimenter (the author of this paper) was present during the whole session. After signing two consent forms, one for the experiment and one for the recording, the subjects were handed a five-page questionnaire in German, one page per object.

Under the heading *Naming unknown objects* each page was divided into three sections corresponding to three phases:

1. In Phase 1 (non-interactive; 3 minutes) individual pools were created.

   Instruction: *Think up three names for this object and write them down.*

2. In Phase 2 (interactive; 6 minutes) the individual pools were joined. The resulting items, up to fifteen, constituted the joined pool from which the subjects had to select in phase 3.

   Instruction: *Share your suggestions with the group and tell them why they are good. Write down the suggestions of the others and compare.*

3. In Phase 3 (non-interactive; 1 minute) subjects had three lines each for (a) making an artificial selection and (b) estimate a natural selection from the candidates in the joined pool.

   Instruction: *Write down* (a) *the names you would personally prefer to use and*
   (b) *those terms that in your opinion would best prevail in a larger community.*

[11] Note that Lewis' (1969) concept of convention can be seen as being based on a second-order preference: The preference for shared sign use is preferred over possible personal preferences for the use of a sign.

Each object picture was projected on the lab's whiteboard for ten minutes, divided into three, six and one minute for the first, second and third phase, respectively; then the picture changed.

In order to merge the projected rankings under 3 (b) it was decided to give the entries in the third column single weight, those in the second one double weight and the winners triple weight. Thus a unified ranking resulted with e.g. *Hamming* (not an existing common noun, but the root is that of hammer and the suffix *-ing* very frequently used) getting the weight ten, because two subjects had it in the first position and two in the second one.

The entries resulting from phase 3 (b) served as basis for deriving the features of natural selection, i.e. those properties that separate the winners from the rest: the winners are supposed to show the highest degree of fitness, the runners-up the second highest, and the third-place finishers the third highest.

3.2. Pretest results: five pools resulting from artificial and estimated natural selection

Already the first cursory look at the data showed such a big difference from what has been the result of the preparatory non-interactive data collection that only one conclusion could be drawn: H5, the Initial Word Prediction Hypothesis underestimates the inventiveness at work in name creation, especially when interaction comes into play, to such a dramatic extent that this hypothesis had to be completely discarded.

Here are the top three names for each object according to the estimated natural selection:

O 1:	1. *Hamming*	2. *Trimolin*	3. *Togolino*
O 2:	1. *blue Willi*	2. *Fantaan*	3. *Geometros*
O 3:	1. *Starcopter*	2. *Starnaut* *Wormster*	3. *Burcumber*
O 4:	1. *Bumbleblue*	2. *Piggyback*	3. *Pipester*
O 5:	1. *Tubedang* *Volereur* *Kinderjet* *Poolboard*	2. *Mobile Puff* *Skyfall Shuttle*	3. *Poolfun* *Blue Gun Pro X8*

Table 1

There were three binary ties, one with Object 3 and two with Object 5. The latter was also special because there was a five-fold tie in the winner’s slot. Strikingly, no knowledge of German is required to understand the proposals except for the modifier of the compound *Kinderjet*: *Kind* is 'child' and *-er* is the linking element.

Fortunately for the replacement of H5 by a better hypothesis, the estimated natural selections gave several hints for getting at the real determinants, especially when compared with the artificial selection.

First, items that were too long were estimated to lose out on their competitors. So quantity, less in terms of roots than in terms of syllables, seems to play a key role.

Second, the ontological categories assumed by H5 were much too specific. So a more abstract and general way of accounting for the descriptive aspect of the candidates had to be found.

Third, the by far most important change on the way from H5 to a new hypothesis based on the results of the pretest consisted in turning a problem into an asset: instead of being considered only an impediment for predictability, the unexpected degree of inventiveness was transformed into an additional predictor.

Thus H4, Darwin's Hypothesis mentioned above, has been partially vindicated, except that 'mere novelty and fashion' turned out to belong among the most important causes for the survival of new words, and not among the less important ones, as Darwin surmised.

3.3. Developing a selection predictor: factors of fitness ranking

In order to come up with a more viable hypothesis than H5 the following question had to be asked: what are the features that make the difference between the 21 winners listed above and the remaining 49 names? Here are some examples from the latter pool:

O 1:	*abstrakte Kanone* 'abstract cannon'	*Dreikopftier* 'three head animal'	*blaues Rudikolon* 'blue rudicolon'
O 2:	*blauer Vogelstrauß* 'blue ostrich'	*Bolzenstandgerät* 'bolt foot device'	*Edler Klopapierhalter* 'upscale toilet paper holder'
O 3:	*Wurstkolben* 'sausage piston'	*Stangenrohr* 'rod pipe'	*Sternenraupe* 'star caterpillar'
O 4:	*Garderraupe* 'garder caterpillar'	*Akropferd* 'akro horse'	*Skybike* 'sky bike'
O 5:	*Flug-wiege* 'flight cradle'	*Kinderprise* 'child pinch'	*Schwimmlandschaft Delux* 'swimming pool landscape deluxe'

Table 2

Comparing tables 1 and 2 provided support both for H3, Müller's Hypothesis (the better, the shorter, the easier survive), and for H4, Darwin's Hypothesis (novelty and fashion foster survival), and thus inspired the following proposal for a word prediction model that consists of four categories.

Echoing Grice (Grice 1975: 45), who was echoing Kant (Kant 1781 [1881]: 71), who in turn had borrowed from Aristotle, these categories will be called Quantity, Quality, Relation, and Manner. In the current context they serve to denote four aspects of fitness for the survival of neologisms:

Quantity:	closeness to optimal length
Quality:	number of fitting object features captured

Relation: number of distinctive object features captured

Manner: degree of originality and fashionability

H6 Final Word Prediction Hypothesis

A German name for a novel object is more likely preferred to its competitors

1. the closer its length is to the optimum,
2. the higher the number of object features it codes,
3. the higher the number of distinctive object features it codes,
4. the higher its degree of originality and fashionability.

The next step was one of the most challenging tasks of the project: developing valid operationalizations of these categories.

1. Quantity In order to deal with the second part of Müller's Hypothesis ('the shorter ones win') the number of words, roots, morphemes and syllables in the winners and the losers of the pretest pools for each object were counted and the following means were calculated:

	words	roots	morphemes	syllables
winners	1.3	1.8	2.15	2.9
losers	1.34	2.28	2.8	3.86
difference	0.040	0.480	0.650	0.969

Table 3

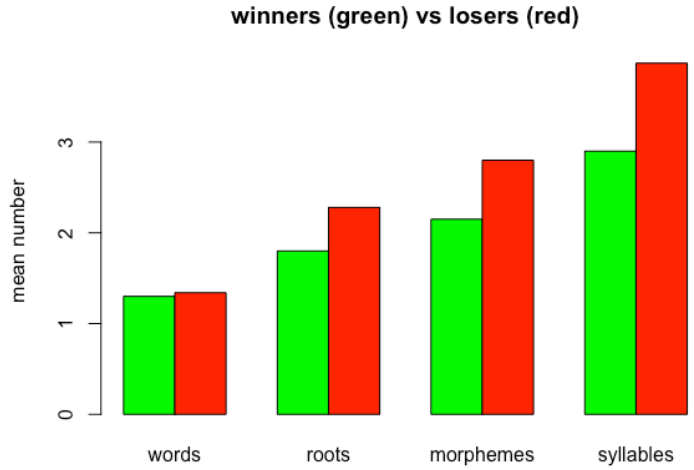


Figure 3

This confirmed part two of Müller's Hypothesis in all possible specifications, but since the difference was most pronounced among syllables the Quantity aspect was operationalized as number of syllables.

However, the preference for shorter names is of course not without limitations, else monosyllabic names would be always preferred. This may be true to some extent for English words (see below), but in the German pretest pools it certainly does not hold. Figure 4 is a chart of the distribution of length in terms of syllable count among the 70 items in the five pools together:

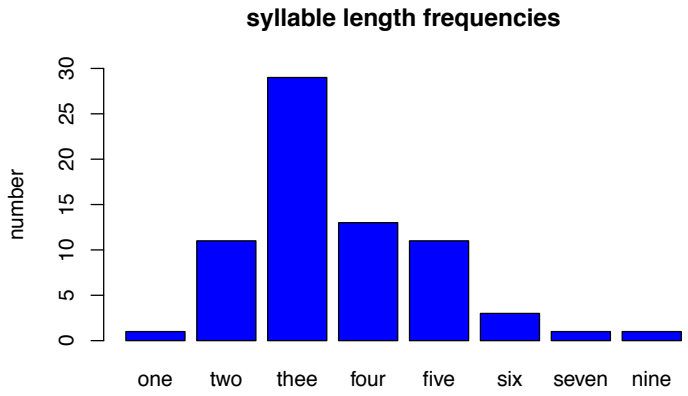


Figure 4

Hence the decision was made against taking the quantity aspect of Müller's Hypothesis literally and in favor of interpreting it as a preference for the optimal number of syllables, which according to the sample is three.

2. Quality The third part of Müller's Hypothesis ('the easier ones win') was much more difficult to operationalize. It was decided to interpret 'easy' as 'easy to remember' or in other words degree of motivation: for instance, when a star is part of the object, names with 'star' are well-motivated and easy to remember, in short, they fit this feature of the object. The more features a name fits, literally or figuratively, the higher its degree of Quality.

3. Relation The most difficult to operationalize was the first part of Müller's Hypothesis ('the better ones win'). Since 'good' can mean anything that increases fitness, it seemed plausible to subsume under Relation the degree to which a name is helpful in discriminating between the five objects. As all objects are blue, the adjective in *blauer Vogelstrauß* 'blue ostrich' fits the color of the object, but it is not distinctive.[12] By contrast, *Three Feet Bird* is distinctive because only object 2 stands on something similar to three feet.

4. Manner According to Darwin's Hypothesis there is a fourth fitness increasing factor, in addition to the three parts of Müller's Hypothesis, namely 'mere novelty and fashion.' The chosen operationalization incorporates both originality ('novelty') and number of foreignisms ('fashion') such as anglicisms or italianisms.

The diversity of word formation devices was astonishing. The pool with its 70 elements contained about 37% German nominal compounds such as *Umschaltknopfgerät* 'toggle button

[12] Interestingly, but not unexpectedly this noun phrase occurs as a name for both object 1 and object 2; it was suggested, of course, by different subjects.

device' or *Formentier* 'forms animal', 30% foreignisms, most of them anglicisms such as *Chick-a-brick* or *Piggyback*, but also italianisms like *Togolino* and one frenchism: *Volereur*. Quite a few suggested names could not be assigned to any current word formation type, among them *Heta*, *Trimolin*, *Rudikolon*, *Fantaan*, and *Flovotov*. All in all, the degree of lexical heterogeneity was too high to allow for an account of word formation types beyond the very superficial observations just stated.

Coming back to the topic of categorization and its relation to naming, especially to the question what insights into category formation can be gained from an analysis of name building, it is obvious that the first and the last categories, the formal factors Quantity and Manner, are unhelpful. What is more, the comparison of the items in table 1 (winners) and table 2 (losers) shows a strong tendency towards preferring short and original names over more transparent, in the sense of better motivated and distinctive, ones. And indeed there is a built-in tradeoff between the formal and the functional categories: the more features are literally or figuratively captured, the higher the probability that the Quantity restriction is exceeded. In sum, the selection process that makes the difference between those expression that stick, and the others that don't tends to decrease the degree to which naming processes are helpful in learning about the categorization procedures behind them.

This is one lesson to be learned from the pretest data: if one is interested in the category formation behind the naming, it is more informative to look at the fleeting stars among the neologies, those that disappear immediately after their creation, whenever they are available, than to analyze those expressions that make it into the lexicon.

The other lesson is to do with the general ontological categories of artifact versus living creature. Interestingly, even though the Fribbles are often described as animal-like creatures the suggested animal names like *Kugelfant* 'sphere (ele)phant' are outnumbered by expressions that imply a conceptualization as artifact. This is interesting because, as mentioned above, artifacts are best characterized in terms of their function and not their visual appearance, whereas the Fribbles stimuli only provide visual information such that possible functions have to be guessed from the appearance.

The question whether this can be interpreted as confirming a human tendency towards conceptualizing the environment in terms of affordances or potential functions must be left to a different occasion.

3.4. Assigning degrees of fitness: the coding algorithm

In order to come up with an algorithm that enables different coders to consistently calculate fitness scores, a further working hypothesis was required: in line with the principle of indifference the four categories were given equal weight, i.e. they were assumed to contribute equally to the overall degree of fitness, with an approximate maximum of 10 in each category. Regarding Quantity and Manner this is easy to achieve by setting the maximum to 10 and 8 + *x*, respectively. For Quality and Relation things are different, since in principle there is no upper limit. But as mentioned above there is a trade-off between these categories and Quantity: it is difficult to increase the number of fitting elements without increasing the

number of syllables. Therefore the algorithm was designed in a way that practically excluded Quality and Relation scores higher than 10, such that the overall maximal score was close to 40.

The complete algorithm for calculating fitness scores is defined in table 4 and the following formulae.

1. Quantity Motivated by the syllable count frequencies shown in figure 4 the coding scheme assigns the maximum of 10 points to words with 3 syllables and decreasing scores with shorter and longer items:

Syllable number	1	2	3	4	5	6	7	8	9
Quantity score	6	8	10	8	6	4	2	1	0

Table 4

2. Quality Here the number of literally fitting aspects *lif* is given twice the weight of the number of figuratively fitting aspects *fif* and the result is doubled. The formula is this:

Quality score := (*lif* * 2 + *fif*) * 2

3. Relation The Relation score can at most equal the Quality score. The formula is the same except that the number of distinctive literally fitting aspects *dlif* and the number of distinctive figuratively fitting aspects *dlif* take the place of *lif* and *fif*, respectively.

Relation score := (*dlif* * 2 + *dfif*) * 2

4. Manner The first factor of the Relation score is the most subjective part of the whole algorithm in asking the coder for an *edo*, an estimate of the degree of originality (Darwin's 'novelty') on a scale ranging from 0 to 8. The second factor *nfe* simply counts the number of foreign elements. So the formula is:

Manner score := *edo* + *nfe*

The overall fitness score formula is simply the sum of the four category scores:

Fitness score := Quantity score + Quality score + Relation score + Manner score

4. Experiment: testing the fitness rank predictor via forced natural selection

4.1. Experiment procedure

The procedure was the same as in the pretest (cf. section 3.1. above), except that this time six new subjects were recruited, four females and two males, and that the five-page questionnaire they were given had been modified in several partly crucial respects.

After the title *Naming unknown objects* each page was again divided into three sections corresponding to three phases (*n* is a number ranging from 1 to 5):

Phase 1: Thinking up suggestions (non-interactive; 2:30 minutes)

Instruction: *How would you want to talk about object n? Come up with three names for this object and write them down.*

Phase 2: Comparing proposals (interactive; 5 minutes)

Instruction: *Share your suggestions with the group and tell them why they are appropriate for communicating about Object n. Write down the suggestions of the others and compare.*

Phase 3: Finding a consensus proposal (interactive; 2:30 minutes)

Instruction: *Write down in column (a) the three top names you would recommend to a larger community for communicating about Object n. Compare and discuss the proposals until you find a consensus. Write down the consensus recommendation in column (b).*

Each object picture was projected on the lab's whiteboard for ten minutes altogether, two and a half minutes for the first, five minutes for the second and again two and a half minutes for the third phase; then the picture changed.

Resulting thus from a process of forced natural selection, the entries obtained in phase 3 section (b) were the ones to be used for testing the quality of the fitness rankings developed on the basis of the pretest data.

4.2. Comparing observed and predicted rankings

4.2.1. Object 1 experiment ranking

For the first object only two of the six subjects came up with the full range of three proposals, three provided two and one participant a single one, such that there were 13 names in 18 cells (72%). The group reached a consensus only for the first rank, the other ones were left open. The winner was *Walking Cam*. The fitness scores provided by the algorithm presented in section 3.4. above for the thirteen names are shown in table 5 and range from 8 to 28, yielding ten ranks (there are three ties). And the item with rank one is *Walking Cam*: perfect match!

fitness score	28	24	23	22	18	18	15	15	14	13	10	10	8
predicted rank	1	2	3	4	5	5	6	6	7	8	9	9	10
observed rank	1	–	–	–	–	–	–	–	–	–	–	–	–
difference	0	–	–	–	–	–	–	–	–	–	–	–	–

Table 5

4.2.2. Object 2 experiment ranking

The second object was seemingly harder to deal with than the first one as witnessed by the records: only 11 of the 18 cells were filled (61%) and the final listings by the six subjects of three top candidates, where they were noted, showed some disagreement regarding the order. Three subjects assumed the consensus list to be 1. *Geometrievogel* 'geometry bird,' 2. *Space-Kugler* 'space blowfish' and 3. *Rundständer* 'round pedestal,' but one put *Space-Kugler* in first position and no other names in second and third place, whereas the other two participants left

all three positions open. Therefore the algorithm used in drawing conclusions from the pretest (end of section 3.1. above) was reactivated and triple, double and single weight was assigned to the first, second and third position, respectively. The result is a tie in the first rank (*Geometrievogel*, *Space-Kugler*), one second ranking word (*Rundständer*) and no third rank.

As documented in table 6, there is a perfect match of the predicted and observed ranks for *Space-Kugler* and *Rundständer*. The predicted fourth rank for *Geometrievogel* underrates the observed first rank by 3.

fitness score	25	25	24	22	22	20	20	17	12	12	6
predicted rank	1	1	2	3	3	4	4	5	6	6	7
observed rank	–	1	2	–	–	–	1	–	–	–	–
difference	–	0	0	–	–	–	–3	–	–	–	–

Table 6

4.2.3. Object 3 experiment ranking

Data harvest was much more bountiful for the third object: its picture elicited 16 out of the 18 possible proposals (89%) and the consensus ranking was unanimous.

The fitness scores listed in table 7 show an unusually gentle slope: They predict two tied winners and seven tied runners-up. The observed ranks, however, are quite different, and there is no perfect match at all. The winner *Sternentaucher* 'star diver' is only minimally underrated by the prediction, the third-ranking *Schnuppenstern* 'snuffing star' is overrated by two ranks, but *Galaxieplattform* 'galaxy platform,' that finished second, is rather strongly underrated, namely by 5.

fitness score	28	28	26	26	26	26	26	26	26	24	23	22	20	16	10	6
predicted rank	1	1	2	2	2	2	2	2	2	3	4	5	6	7	8	9
observed rank	3	–	1	–	–	–	–	–	–	–	–	–	–	2	–	–
difference	+2	–	–1	–	–	–	–	–	–	–	–	–	–	–5	–	–

Table 7

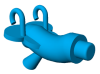
4.2.4. Object 4 experiment ranking

The fourth object was apparently less inspiring for the group: they produced only 13 out of 18 (72%) possible names. Except for one dissenting vote the consensus ranking was as follows: 1. *Breakdancer*, 2. *Insektenständer* 'insect stand' and 3. *Krauler* 'crawler.'

Table 8 shows that observed and predicted ranks again form a perfect match, as with object 1, but this time with three pairs.

fitness score	26	25	24	22	20	17	16	14	14	12	12	10	8
predicted rank	1	2	3	4	5	6	7	8	8	9	9	10	11
observed rank	1	2	3	–	–	–	–	–	–	–	–	–	–
difference	0	0	0	–	–	–	–	–	–	–	–	–	–

Table 8



4.2.5. Object 5 experiment ranking

Even fewer names were proposed for object 5, namely only 12 out of 18 (66%). The group agreed (again with one exception) on the following top candidates: 1. *Schlauchdüse* 'hose nozzle,' 2. *Stöpsler* 'plugger' and 3. *Tubenelefant* 'tube elephant.' Due to the dissenting vote the latter rose to second rank, yielding a tie with *Stöpsler*.

The fitness scores shown in table 9 predict three tied winners: *Schwimmbadtube* 'swimming pool tube' and *Henkelröhre* 'handle pipe' in addition to *Schlauchdüse.* Only the latter matches the observed winner in a perfect match. Regarding the tied second positions the prediction hit the mark with *Stöpsler* and underrated slightly (by 1) *Tubenelefant*.

fitness score	28	28	28	22	21	20	20	18	17	14	13	12
predicted rank	1	1	1	2	3	4	4	5	6	7	8	9
observed rank	–	–	1	2	2	–	–	–	–	–	–	–
difference	–	–	0	0	–1	–	–	–	–	–	–	–

Table 9

4.2.6. Overview of experiment rankings

Table 10 shows the main results of the experiment: given the total number of 13 pairs for comparison, and counting over- and underrating score differences of 1 as near matches and bigger differences as strong mismatches, there are 8 full matches (62%) and 10 full or near matches (77%), corresponding to 5 mismatches (39%) and only 3 strong mismatches (23%). The average mismatch value was a 0.62 underrating score.

Labels by object Experim nat sel	*Syllables*	*Quantity*	*Quality*	*Relation*	*Manner*	*Fitness*	*E pred rank*	*E obs rank*	*Match*
Walking Cam	3	10	4	4	10	28	1	1	**MATCH**
Geometrie-Vogel	6	4	6	4	6	20	4	1	u_3
Space-Kugler	3	10	6	4	5	25	1	1	**MATCH**
Rundständer	3	10	8	4	2	24	2	2	**MATCH**
Sternentaucher	4	8	6	6	6	26	2	1	u_1
Galaxie-Plattform	5	6	2	0	8	16	7	2	u_5
Schnuppenstern	3	10	6	4	8	28	1	3	o_2
Breakdancer	3	8	4	4	10	26	**1**	**1**	**MATCH**
Insektenständer	5	6	6	6	7	25	2	2	**MATCH**
Krauler	2	8	4	4	8	24	3	3	**MATCH**
Schlauchdüse	3	10	6	6	6	28	1	1	**MATCH**
Stöpsler	2	8	4	4	6	22	2	2	**MATCH**
Tubenelefant	5	6	6	4	5	21	3	2	u_1
AVERAGE Exper	3.6153846	8	5.23071	4.15384	6.69230	24.0769			u_0,62
STDEV.P Exper	1.2113858	1.9215378	1.47564	1.45951	2.08970	3.40726			

Table 10

In view of these results it seemed justified to state that the four fitness categories derived from the pretest performed surprisingly well.

4.3. Discussion of the experiment results

4.3.1. Limits of ecological validity

The experiment and its outcome left of course many questions open. It was clear right from the beginning that the question of how close the lab setting may come to situations of everyday life was an open issue.[13] As a fallback position (position 1) one could hold the view that at least regarding small group decision processes where a pool of candidates is generated the ecological validity is sufficiently good. But this may also be an overly restrictive option.

The following real-life examples are selected to shed some light on this issue.

Paradise Papers

A more recent case than the Engelbart story and one that by contrast did not meet the small group restriction was the christening process for the set of 13.4 million confidential electronic documents about offshore investments leaked in 2017 to Frederik Obermaier and Bastian Obermayer, two German investigative journalists at the newspaper *Süddeutsche Zeitung*, who shared them with an international network of more than 380 journalists. Here is an excerpt of what they wrote (Obermaier and Obermayer 2017; translation DZ):

[13] At the Bologna meeting the present volume is based on Mira Ariel gave voice to a rather strong skepticism about the transferability of the results beyond the case of Engelbart and his team.

"**How the Paradise Papers originated**

Answers to the most important questions regarding the new data leaked to the SZ

Whence the name Paradise Papers?

This time, unlike the case of the Panama Papers, it was not one country that played the central role, but many countries. Since these are all so-called *Steuerparadiese* ('tax havens'), the name Paradise Papers was born."

This excited the curiosity of the author and so he sent Frederik Obermaier an email asking for the way the consensus for the final name Paradise Papers was reached. He kindly and immediately answered as follows (translation DZ):

"The naming process involved the entire international Paradise Papers team, more than 300 journalists from around the world. There were a number of aspects that played a role in the discussion, including:

- Does it reflect the core of the research?

- Is the name also understandable in other languages, such as French or Spanish, or is it easy to translate?

- Is alliteration possible? This has proven itself in the past (Panama Papers)

- Does the name also work in social media, for example in conjunction with a hashtag? – ParadisePapers (according to experience, it is often incomprehensible for more than two words)

In my view, "Paradise Papers" has prevailed, as this project affected several countries and therefore a limitation to individual countries would not have been fair to the project. Many partners liked the analogy with the Panama Papers and the Pentagon Papers as well as the alliteration."

Regrettably, the follow-up request for other members of the pool from which the winner had been selected was rejected; in view of future occasions the journalist would rather not disclose them.

Still the newspaper case can be seen as encouraging a stronger position (position 2) to the effect that the ecological validity of the approach is sufficiently good at least regarding group processes of selection from a pool of candidates, dropping thus the small group requirement.

Brexit

It also has been argued that the prior emergence of a candidate pool from which the neology is selected may be more the exception that the rule. Here is another real-live example that relates to this issue.

On 25 December 2016 Tom Moseley, a political reporter at BBC wrote an article *The rise of the word Brexit*. Some passages are interesting in the present context (Moseley 2016):

"Who coined the phrase?

The Oxford English Dictionary awarded this honour to Peter Wilding when it added Brexit to its volumes recently.

Mr Wilding is the founder and director of the British Influence think tank - and campaigned for the UK to Remain in the EU in June's referendum.

He wrote about "Brexit" in May 2012, eight months before the then Prime Minister David Cameron had announced he would be holding a referendum.

> "Unless a clear view is pushed that Britain must lead in Europe at the very least to achieve the completion of the single market then the portmanteau for Greek euro exit might be followed by another sad word, Brexit," he predicted. […]
>
> Brexit not Brixit
>
> It could have all been (slightly) different.
>
> Brexit was far from set in stone, and faced early competition with an alternative version featuring the following month in an Economist article predicting that "a Brixit looms for several reasons".
>
> In August 2012, investment bank Nomura made waves when it warned the City in a report that a "Brixit" was "increasingly likely", while the same term was used in a Daily Mail column urging: "Bring on the 'Brixit'."
>
> But Brexit prevailed, although it was another three years before its use really took off."

Even though this is only anecdotal evidence, it hints at the possible tenability of a third position that is still one grade stronger than its predecessors. The seeming absence of an overt pool of candidates may often result from the fact that competing expressions are forgotten, as in the case just mentioned. And even if there is neither a group nor an overt pool, the assumption that in word production there is always a selection among competing alternatives at least in the mind is rather well supported by current neuroscience (cf. Hagoort et al. 2009).

These considerations indicate that the more cautious views of position 1 and position 2 may be safely given up, reaching the admittedly somewhat bold conclusion that neither the restriction to small group size nor to situations with an overt candidate pool is appropriate and that the ecological validity of the approach is possibly as general as can be (position 3). There are, however, other constraints that will be discussed in the following section.

4.3.2. Other possible constraints

As already mentioned in section 3.3. above the fitness score for item length in terms of syllable count is not universal. It is rather highly plausible that the finding of a preference for trisyllabicity in new words of German does not necessarily carry over to other languages. Here is a real-life example that points in this direction.

Cell phone vs. mobile phone

Preliminary evidence for a possible language specificity of optimal item length comes from comparing the proportions of *cell phone* (two syllables) and *mobile phone* (three syllables) in American and in British English, respectively, as documented in Google's Ngram viewer: whereas British English shows a strong preference for the longer over the shorter term (figure 5), the situation in American English is exactly the opposite (figure 6).

British English

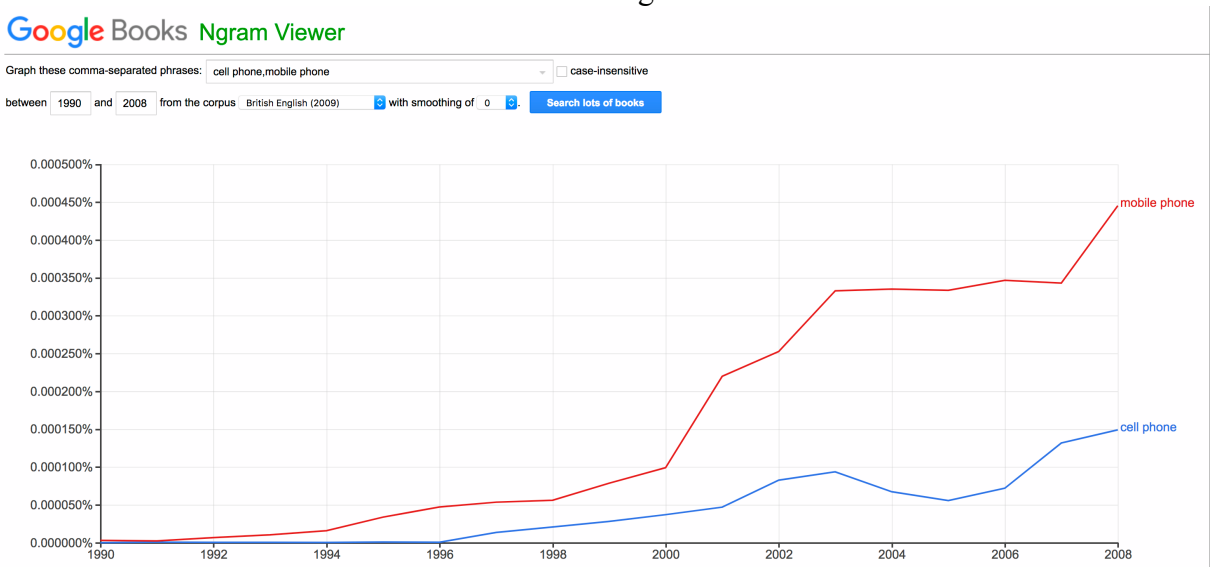


Figure 5

American English

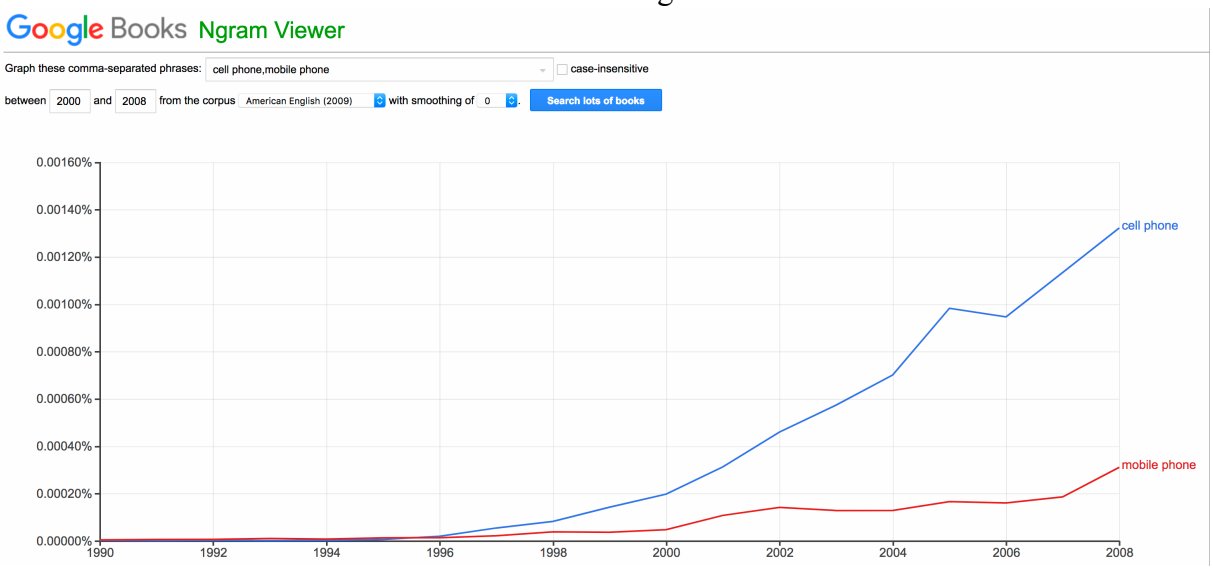


Figure 6

Moreover, it seems reasonable to assume that not only quantity, but also the other three fitness categories may vary from language to language and from culture to culture requiring thus an adequate parameterization.

The high proportion of words with foreignisms (23%) among the 13 winners of the experiment, for instance, is probably among other things due to the age of the subjects: most of them were in their twenties. Hence age may be another relevant parameter.

Not only cultural differences, including those based on membership in different generations, play a role in parameter setting for the fitness categories, but also individual preferences and

distinct kinds of group dynamics, as shown by the different degrees of participant productivity in the pretest and in the experiment. And as pointed out by Frederik Obermaier, the internet-based exchange of about 380 journalists constituted an intriguing kind of consensus finding where an even richer spectrum of fitness criteria seemed to play a role.

4.3.3. Tackling the replicability issue

The so-called replication crisis does not only affect psychological science (Pashler and Wagenmakers 2012), but the entire range of empirical humanities including linguistics (Peels 2019). This situation and the fact that the four fitness categories derived from the pretest performed possibly a little too well in the experiment were the main reasons for postponing the publication of its results until a replication had been performed. This is the topic of the following section.

5. Replication: retesting the fitness rank predictor

5.1. Replication procedure

In order to control for possible hidden confounds it was the idea of the replication to keep the procedure as close as possible to the experiment procedure described above in section 4.1. It took place in the same MELESSA lab as the pretest and the initial experiment, five months after the latter. Six new subjects were recruited through the same university service, this time three females and three males, and they were given exactly the same questionnaire and instructions as the participants in the experiment.

5.2. Comparing observed and predicted replication rankings

5.2.1. Object 1 replication ranking

Displaying a remarkable degree of creativity the replication team invented 18 different names for the 18 available slots. However, their consensus ranking left the third place empty. This time the prediction based on the fitness scores missed the observed ranking dramatically: the winner was *Elefantenwasserhahn* 'elephant water tap' and the runner-up *Geometrino*, the former was underrated by 5 and the latter even by 6 (table 11).

fitness score	28	28	26	26	23	20	20	19	18	18	18	17	15	15	14	13	8	5
predicted rank	1	1	2	2	3	4	4	5	6	6	6	7	8	8	9	10	11	12
observed rank	–	–	–	–	–	–	–	–	1	–	–	–	2	–	–	–	–	–
difference	–	–	–	–	–	–	–	–	–5	–	–	–	–6	–	–	–	–	–

Table 11

5.2.2. Object 2 replication ranking

By contrast with the experiment group, the replication team members didn't have any problems with finding 18 different names for the second object as well. There was one deviation in the consensus on the top three items (first and second rank were switched by one

participant), but without consequences for the outcome. The winner and runners-up were *Chippihocker* 'chippy stool', *Kugelkamera* 'spherical camera' and *Einrichtungsstrauß* 'furniture ostrich/bouquet,' respectively, and the prediction missed the latter two just by one. However, the winner was rather strongly underestimated, namely by 5 (table 12).

fitness score	26	25	24	24	24	23	23	22	22	22	22	21	21	20	20	16	13	12
predicted rank	1	2	3	3	3	4	4	5	5	5	5	6	6	7	7	8	9	10
observed rank	–	–	–	2	–	3	–	–	–	–	–	1	–	–	–	–	–	–
difference	–	–	–	–1	–	–1	–	–	–	–	–	–5	–	–	–	–	–	–

Table 12

5.2.3. Object 3 replication ranking

As with the first two objects the replication team came up with the maximum of 18 different names, and this time the consensus was faultless: the top three proposals were *Fastfoodstern* 'fast food star', *Gourmetwurst* 'gourmet sausage' and *Sternenkrabbler* 'star crawler,' in this order. Fitness scores predicted the winner exactly and missed the third-place finisher just by 1, whereas the second-ranking item was grossly (by 4) underrated (table 13).

fitness score	32	30	30	28	28	27	26	26	26	25	24	22	22	22	21	21	20	20
predicted rank	1	2	2	3	3	4	5	5	5	6	7	8	8	8	9	9	10	10
observed rank	1	3	–	–	–	–	–	–	–	2	–	–	–	–	–	–	–	–
difference	0	+1	–	–	–	–	–	–	–	–4	–	–	–	–	–	–	–	–

Table 13

5.2.4. Object 4 replication ranking

Once more 100% of the provided cells were filled with 18 different names and a flawless consensus was achieved with the following order of finishers: *Wurstventilator* 'sausage fan,' *Kabelhalter* 'cable holder' and *Hakenprothese* 'hook prosthesis.' The fitness scores predicted a ranking that matched this outcome perfectly for the second and third position, the winner, however, was substantially (by 5) underrated (table 14).

fitness score	30	26	26	24	24	23	22	22	21	21	20	20	19	18	18	17	16	14
predicted rank	1	2	2	3	3	4	5	5	6	6	7	7	8	9	9	10	11	12
observed rank	–	2	–	3	–	–	–	–	1	–	–	–	–	–	–	–	–	–
difference	–	0	–	0	–	–	–	–	–5	–	–	–	–	–	–	–	–	–

Table 14

5.2.5. Object 5 replication ranking

This was the only object for which one member of the highly creative replication group found only two names, reducing thus the completeness score to 99%. The consensus again was unanimous and these were the three top ranking proposals: *Zahnpastapistole* 'toothpaste pistol,' *Liegerollator* 'recumbent walker' and *Knopfaufhänger* 'button hanger.' Here, the performance of the predictor was rather poor: the best-fitting score was the one for *Knopfaufhänger*, which was overrated by 2, the winner *Zahnpastapistole* was underrated by 3, and the underrating score of 7 for the second-ranking *Liegerollator* was the strongest mismatch occurring in experiment and replication together (table 15).

fitness score	30	30	28	27	26	26	24	24	23	22	21	18	18	18	18	17	16
predicted rank	1	1	2	3	4	4	5	5	6	7	8	9	9	9	9	10	11
observed rank	3	–	–	–	1	–	–	–	–	–	–	2	–	–	–	–	–
difference	+2	–	–	–	–3	–	–	–	–	–	–	–7	–	–	–	–	–

Table 15

5.2.6. Overview of replication rankings

The extract from the replication results shown in table 16 looks much less promising than its experiment counterpart in table 10 above: given the total number of 14 pairs of predicted and observed ranks and counting again over- and underrating score differences of 1 as near matches and larger differences as strong mismatches, the number of full matches dropped from 8 (62%) to 3 (21%) and the number of full or near matches fell from 10 (77%) to 7 (50%); in other words the number of mismatches increased dramatically from 5 (39%) to 11 (79%), among them 7 (50%) strong mismatches as opposed to only 3 (23%). Finally, the average mismatch value has almost quadrupled from a 0.62 to a 2.43 underrating score.

Labels by object Replica nat sel	*Syllable*	*Quantity*	*Quality*	*Relation*	*Manner*	*Fitness*	*R pre rank*	*R obs rank*	*Match*
Elefantenwasserhahn	7	2	4	4	8	18	6	1	u_5
Geometrino	5	6	2	0	7	15	8	2	u_6
Chippihocker	4	8	2	0	10	20	6	1	u_5
Kugelkamera	5	6	6	6	6	24	3	2	u_1
Einrichtungsstrauß	4	8	2	2	7	19	4	3	u_1
Fastfoodstern	3	10	6	6	8	30	**1**	**1**	**MATCH**
Gourmetwurst	3	10	4	4	9	27	6	2	u_4
Sternenkrabbler	4	8	6	6	7	27	2	3	o_1
Wurstventilator	5	6	4	4	7	21	1	6	under_5
Kabelhalter	4	8	6	6	6	26	**2**	**2**	**MATCH**
Hakenprothese	5	6	4	4	8	24	**3**	**3**	**MATCH**
Zahnpastapistole	6	4	8	8	8	28	4	1	u_3
Liegerollator	5	6	4	4	8	22	9	2	u_7
Knopfaufhänger	4	8	6	6	5	25	1	3	o_2
AVERAGE Replic	4.57142	6.85714	4.57142	4.28571	7.42857	23.285			u_2.43
STDEV.P Replic	1.04978	2.09956	1.76126	2.2497	1.23717	4.1305			

Table 16

Understandably these figures suffice to curb any enthusiasm about the performance of the fitness score as valid predictor for neologisms. However, since the replication results could have been worse, it was decided to run a less superficial statistical analysis, which will be presented below in section 6.1., after some remarks on the replication.

5.3. Discussion of the replication

5.3.1. Group dynamics

Why was there such a strong difference in the productivity of the experiment group compared to the replication group? Both had 90 cells to fill, but the first group filled less than three out of four cells (72%), whereas the second one missed the 100% score only by one. (For comparison, the productivity of the pretest group was 93%.)

The observations of the experimenter, who was present during all three sessions support the plausibility of an explanation in terms of a pronounced difference in the group dynamics: the replication group was by far the liveliest team. Right from the beginning a male subject took the lead and moderated efficiently the consensus finding process that progressed somewhat laboriously in the experiment group. In addition, there was also a remarkably good spirit, everybody enjoyed the game and was having fun. By contrast, members of the experiment group were relatively shy and proceeded much more cautiously.

Assuming that in general groups are more similar to the experiment team than to that of the replication could vindicate the performance of the prediction tool to some extent.

5.3.2. Optimizing fitness categories?

Another consideration that suggested itself as a consequence of comparing the outcomes of experiment and replication was the option of fine-tuning the fitness score algorithm.

Taking the category of quantity as an example, another look at figure 4 above, charting the distribution of the 70 items in the pretest pool according to their length in terms of syllable counts shows that the difference between the values for four and five syllables is rather small. This could motivate replacing the coding scheme of table 4, where the decline of the score with an increasing number of syllables after the optimum is 10 – 8 – 6 – 4, by the sequence 10 – 8 – 8 – 4 as in table 17:

Syllable number	1	2	3	4	5	6	7	8	9
Quantity score	6	8	10	8	8	4	2	1	0

Table 17

As a result the number of full matches in the replication would decrease from 3 (21%) to 2 (14%), and that number of full or near matches from 7 (50%) to 6 (42%), with the corresponding increases in the numbers of (strong) mismatches. On the other hand the average mismatch value would decrease considerably, by about 30% (from an underrating of 2.43 to 1.71), so if the average mismatch value is considered as most important this change in the fitness score algorithm could indeed slightly improve the performance.

However, since the advantage of this modification seems rather slim, it was not taken into account by the statistics presented below, which are based on the original unmodified schema.

6. Overall analysis

6.1. Ways of estimating correlation strength of incomplete rankings with ties

The primary problem for a statistical analysis posed by our data can be easily read off a sample data set. Table 13, repeated here for convenience as table 18, shows the fitness scores of the names proposed for object 3 in the replication together with the resulting predicted ranking, the observed ranking and the differences, showing one perfect match, one overranking by 1 and one underranking by 4.

fitness score	32	30	30	28	28	27	26	26	26	25	24	22	22	22	21	21	20	20
predicted rank	1	2	2	3	3	4	5	5	5	6	7	8	8	8	9	9	10	10
observed rank	1	3	–	–	–	–	–	–	–	2	–	–	–	–	–	–	–	–
difference	0	+1	–	–	–	–	–	–	–	–4	–	–	–	–	–	–	–	–

Table 18

The considerable number of ties (four double and two triple ties resulting in a rather smooth decline of only 12 points over 18 items) is not a big problem, but the main challenge comes from the high percentage of missing data: Out of 36 data cells pairing predicted with observed ranks only 20 are filled. A simple way of dealing with this situation would of course be a

radical cut-down on the data by throwing away all incomplete pairs of predicted and observed ranks, keeping only 6 of the 36 data points.

An application of Kendall's tau-b rank correlation coefficient (Kendall 1945),[14] ranging from -1 to +1, to all ten data sets yields the following numerical (table 18) and graphical (figure 7) results. In the figure the left column of each column pair represents the experiment and the right, slightly darker one the replication:

Minimal results

Experiment	Object 1 e	Object 2 e	Object 3 e	Object 4 e	Object 5 e	MEAN
tau-b	1	0	-0.2	1	0.82	0.524
p95greater	0	0.5	0.65	0.17	0.11	0.286
Replication	**Object 1 r**	**Object 2 r**	**Object 3 r**	**Object 4 r**	**Object 5 r**	
tau-b	1	-0.333	0.5	-0.333	-0.333	0.100
p95greater	0.5	0.833	0.333	0.833	0.833	0.666
tau-b total		0.319				
p95greater total		0.512				

Table 19

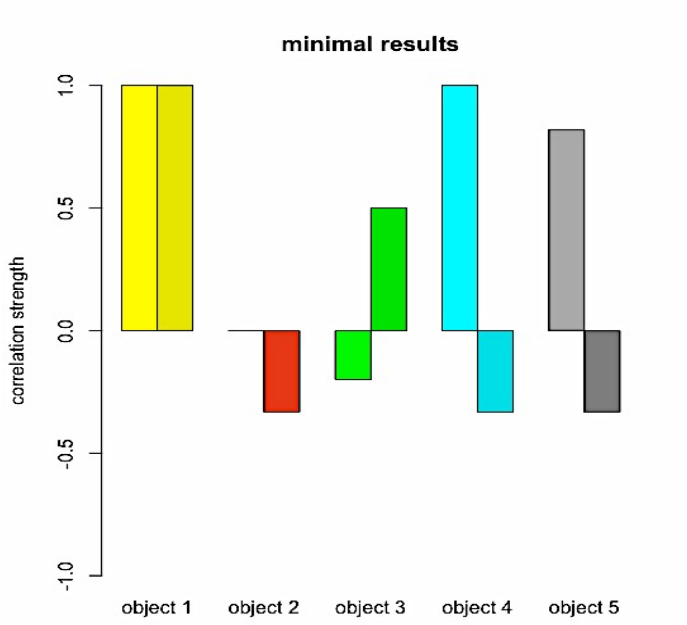


Figure 7

Obviously, disregarding the majority of the collected data results in a notable contrast between the experiment and the replication: except for object 1 they disagree in all cases on the support of the null hypothesis which predicts there to be no or even a negative correlation between predicted and observed rankings. Given that in view of the small number of data

[14] The main reason for choosing this algorithm is its being designed for dealing with ties, which occur frequently in the data of this study.

points the *p*-values are expectably poor, the quality of the predictor is mainly reflected in the strength of a positive correlation, which on average is much better in the experiment's data set than in that of the replication.

But the predicted rank values for those items that lack an observed rank number contain information that does not have to be entirely discarded: the missing values are not completely random, but rather strongly constrained by the requirement of constituting a ranking as well. Unfortunately, although there is an impressively rich literature on missing data,[15] nothing seemed to fit exactly the rank correlation patterns at stake in the present data.

Therefore a specific custom-tailored algorithm had to be developed. In order to get all the missings in one sequence the tables were reordered according to the observed ranking, transforming thus for instance table 18 into table 20 by rearranging the three complete data pairs:

fitness score	32	25	30	30	28	28	27	26	26	26	24	22	22	22	21	21	20	20
predicted rank	1	6	2	2	3	3	4	5	5	5	7	8	8	8	9	9	10	10
observed rank	1	2	3	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
difference	0	–4	+1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–

Table 20

Then a somewhat more sophisticated way of dealing with the missing data than the radical cut leading to the minimal results consists in taking the average of a sufficiently high number of random samples (vectors of length 15 in the sample case) used to fill the gaps.

That could be adequate if there were no further constraints on those randomly chosen vectors apart from length. But as mentioned above there are further restrictions: since ties are possible, the first vector element has to be either 3 again or 4, and the last vector element some integer between 3 and 18. In other words, the condition on the sequence of numbers requires a continuous increasing weak order, which reduces the number of options from 15^{15} to 2^{15}, i.e., from a number with eighteen digits to 32,768. As a result the random sampling has to choose only from this much smaller set.

But what happens to the predicted partners of the thus imputed 'observed' ranks? Looking at table 20 may be misleading, since although it has been sorted, except for the first three columns, in terms of the 'predicted rank' row, for the missing data the order of the corresponding predicted ranks and fitness scores has not been changed. But the predicted rank orders that would result from the different completions of the 'observed rank' line are entirely unknown, meaning that with the exception mentioned above table 20 is only one of a enormous number of equivalent arrangements. What is known is the identity and frequency of the rank numbers that possibly have to be rearranged.

[15] As of November 6, 2019, Google Scholar records 10,428 publications quoting the highly influential article 'Missing data: our view of the state of the art' (Schafer and Graham 2002).

Unfortunately, by contrast with the completed observed rank row where the number of possibilities for n missing values is 2^n, the number of permutations of a given predicted rank order rises factorially in the worst case. In the given example the number of possible orderings is 59,875,200, and generating all these permutations exceeds the limits of the computing power of a desktop.

The algorithm that was developed in the light of these considerations[16] in order to satisfy the relevant constraints proceeds in five steps.

First, the set of possible completions of the predicted rank order is generated, which consists of the permutations of the rank positions without an observed partner. In the case of table 20 the permutation input is the single-row table 21 attained from table 20 by clipping off the first three cells of the second line:

2	3	3	4	5	5	5	7	8	8	8	9	9	10	10

Table 21

In view of (a) the potential factorial growth of the number of permutations, mitigated only by the amount of ties, and (b) the computational limits mentioned above vectors exceeding the length of 12 like the one in table 21 are made to fit the ceiling by cutting off the last three (shaded) elements.

Second, the set of possible completions of the observed rank row is calculated. In the present example this is the set of possible continuations (continuous increasing weak orders) of length 12 of the three top ranks.

Third, for control purposes, sample matrices are built that combine instances of the two completed rows. Tables 22a and 22b are two examples.

fitness score	32	25	30												
predicted rank	1	6	2	8	3	2	5	5	7	3	9	4	8	5	8
observed rank	1	2	3	3	3	4	4	4	4	5	5	6	7	7	7
difference	0	–4	+1	–5	0	+2	–1	–1	–3	+2	–4	+2	–1	+2	–1

Table 22a

fitness score	32	25	30												
predicted rank	1	6	2	5	8	8	2	3	4	5	9	7	5	3	8
observed rank	1	2	3	3	3	3	3	3	4	4	4	4	5	5	5
difference	0	–4	+1	–2	–5	–5	+1	–1	0	–1	–4	–3	0	+2	–3

Table 22b

Fourth, a Kendall tau-b correlation test with confidence level 0.95 and the alternative hypothesis of a positive correlation coefficient is run 10,000 times, drawing thus up to 10,000

[16] The initial ideas was inspired by a suggestion from Felix Schönbrodt (p.c.).

different random samples from the set of possible orderings of predicted ranks on the one hand and of possible continuations of the observed ones on the other.

In the present case this repetition procedure made the resulting tau-b and *p*-values stable at least in the first two decimal positions.

Fifth, the procedure is run three times and the means of the results are calculated.

6.2. Statistical overview of experiment and replication results

A comparison of minimal (table 19 and figure 7) and full results (table 23 and figure 8) elucidates the benefit gained by incorporating the (almost) full amount of data on the prediction level and the restricted set of possible continuations on the observation level: the extremes, especially the strong contradictions between experiment and replication, are gone and the number of negative results has dropped from three to one.

Full results

Experiment	Object 1 e	Object 2 e	Object 3 e	Object 4 e	Object 5 e	MEAN
tau-b	0.16	0.26	0.11	**0.44**	0.30	0.254
p95greater	0.30	0.20	0.33	0.05 *	0.15	0.206
Replication	**Object 1 r**	**Object 2 r**	**Object 3 r**	**Object 4 r**	**Object 5 r**	
tau-b	*-0.04*	*0.01*	0.21	0.28	*0.05*	0.102
p95greater	0.56	0.49	0.20	0.14	0.42	0.362

tau-b total	0.178
p95greater total	0.284

Table 23

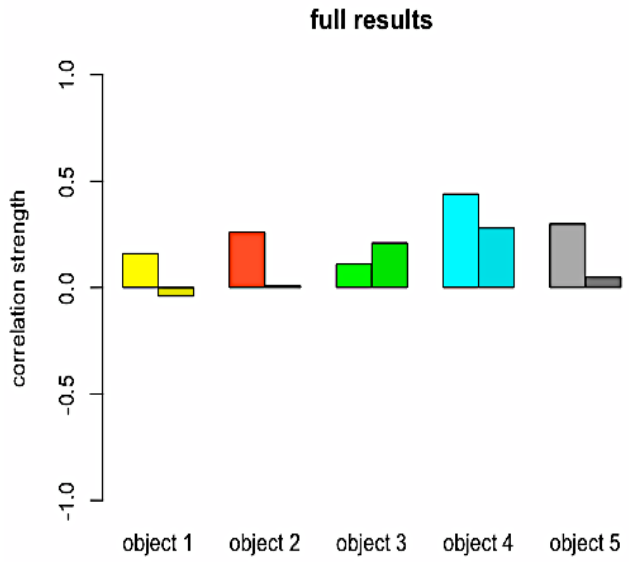


Figure 8

According to Cohen's (1988) frequently used suggestion for evaluating the correlation coefficient as effect size (here correlation strength) estimators, this size is called as follows:

- *small* if the coefficients are between.10 and .29,
- *medium* for coefficients between .30 and .49, and
- *large* in the case of coefficients of .50 and higher.

Using this terminology we can say that among the ten full individual results there is one medium size effect (0.44 for object 4 in the experiment; printed in boldface), six small effects (for the other four objects in the experiment and for two in the replication), and three negligible ones (all in the replication; printed in italics), one of them even negative. The mean of the experiment effect size is in the upper quarter of the range of small (0.254), that of the replication less than half as big, but still in the same category, albeit by a rather tiny margin, and the overall mean is therefore small as well.

The probabilities of the obtained (or worse) results under the assumption of a one-tailed null hypothesis (no or even a negative correlation between predicted and observed rankings) at a .05 significance level (.05 *p*-values) are almost all by far too high, which is what one would expect given the small amount of data. There is one exception, the result for object 4 in the experiment, which is just significant (marked by an asterisk) in that sense. In any case, the *p*-value average of 0.206 for the experiment means that in four out of five cases the null hypothesis fails to predict the obtained results.

Moreover, there is a well-motivated agreement among most researchers that the usefulness of a pilot study cannot be read off standard significance levels like the usual .05-level, the reason being that virtually by definition pilot studies are underpowered in view of this level. Here is a pertinent quote from an article dealing with medical research methodology:

> "We recommend that in pilot trials the focus should be on descriptive statistics and estimation, using confidence intervals, rather than formal hypothesis testing and that confidence intervals other than 95% confidence intervals, such as 85% or 75%, be used for the estimation." (Lee at al. 2014: 1)

In view of the fact that a 95% confidence interval and a 75% confidence interval correspond to a 05% and a 25% significance level, respectively, we can state that five out of the altogether ten results for individual objects are in line with this recommendation, and so does the average of the five results of the initial experiment.

All in all, the data collected in this pilot study provide some preliminary evidence for the hypothesis of a positive correlation, albeit a small one, between the predicted and the observed rankings and as a result for the usefulness of H3 and H4, Müller's and Darwin's Hypothesis in our operationalization, predicting some core features of neologisms.

7. Summary and outlook

In the light of the considerations and results discussed and presented in this paper, what is the answer to the question asked in its title? Is it really the case that new words are predictable? A short, straightforward and rough answer is simply: no.

But there is evidence that the short answer is true only to some extent. Admittedly, predicting the emergence of demand for a new word is possibly beyond of the scope of any theory, let alone linguistics. But regarding their linguistic aspects a longer and less simplifying answer is this: new words are at least not completely unpredictable. There is reason to assume that in line with H6, the Final Word Prediction Hypothesis, the four factors quantity (word length), quality (descriptive fit), relation (discriminativeness) and manner (attractiveness) play crucial roles in their selection.

As outlined above (section 1.3.) this hypothesis is far from being new, it has been formulated in a rather speculative way already one and a half centuries ago by two outstanding scientists of their epoch who mentioned those factors as determining the survival of certain new words at the expense of others: H3, Müller's Hypothesis, comprises the first three of these factors, and H4, Darwin's Hypothesis, states the last one.

What is new is that the present pilot study has provided some experimental support for the hypothesis. The four fitness categories proposed did reasonably, in part even surprisingly well. Even though due to its lack of power this exploration could not come up with really strong evidence, its results are compatible with the notion that a more extensive investigation, possibly with an improved weighting of factors, could lead to a truly predictive theory of core determinants of the creation, selection and survival of new words.

In view of these findings there is reason to assume that Q1, Engelbart’s Question, to wit 'Why do some neologies stick, whereas other don’t?' is not hopelessly hard to answer.

Last, but not least I want to submit that the present study encourages using the concept of natural selection via community preference with both its conscious and its uncontrollable aspects for the investigation of other phenomena, linguistic and non-linguistic, in the domain of cultural evolution, and that it definitely seems to be worth of further exploration.

Acknowledgements

The author gratefully acknowledges the valuable contributions of the following people and institutions (in the order of appearance). Thanks go to

• Daniel Casasanto for lively discussions on the occasion of a joint seminar on metaphor, metonymy and indirect conceptualizations (cf. Casasanto 2009b), and for bringing the Fribbles to my attention, providing me thus with the first spark of inspiration for the present study: the idea of forcing people to produce indirect conceptualizations by presenting them with stimuli for which no pre-existing direct conceptualizations are available,

• the Fribbles: stimulus images courtesy of Michael J. Tarr, Center for the Neural Basis of Cognition and Department of Psychology, Carnegie Mellon University, http://www.tarrlab.org/,

• Daphné Kerremans and Hans-Jörg Schmid for stimulating discussions on neologisms,

• Lisa Spantig and the crew of the Munich Experimental Laboratory for Economic and Social Sciences (MELESSA) of Ludwig-Maximilians University for providing laboratory resources,

• the audiences of the presentations of preliminary results on October 19th and December 9th 2017 at the universities of Bologna and Munich,

• Hans-Jörg Schmid for generous financial support (subject payment and lab use fee),

• Helmut Küchenhoff and Felix Schönbrodt for helpful hints regarding statistical methods,

• R: A language and environment for statistical computing (R Core Team 2018) for developing the computational tool for the statistics,

• Randy Lai (Lai 2018) for writing the R package that computed the permutations that were needed,

• two anonymous reviewers for providing truly helpful suggestions for the present revised version of the paper.

8. References

Aronoff, Mark. 2016. "Competition and the lexicon." In: Elia, Annibale and Iacobini, Claudio and Voghera, Miriam (eds.): *Livelli di analisi e fenomeni di interfaccia*. Roma: Bulzoni. 39-52.

Aronoff, Mark. 2017. "Darwinism tested by the science of language." In Claire Bowern, Laurence Horn and Raffaella Zanuttini (eds.), *On looking into words (and beyond)*. 443-456. Berlin: Language Science Press.

Beaumont, Claudine. 2008. "Computer mouse celebrates 40th birthday." https://www.telegraph.co.uk/technology/news/3538800/Computer-mouse-celebrates-40th-birthday.html

Barry TJ, Griffith JW, De Rossi S and Hermans D. 2014. "Meet the Fribbles: novel stimuli for use within behavioural research." *Frontiers in Psychology*, *5*, 103. doi: 10.3389/fpsyg.2014.00103

Bloomfield, Leonard. 1926. "A set of postulates for the science of language." *Language* 2.3: 153-164.

Casasanto, Daniel. 2009a. "Embodiment of abstract concepts: good and bad in right-and left-handers." *Journal of Experimental Psychology: General* 138.3: 351.

Casasanto, Daniel. 2009b. "When is a linguistic metaphor a conceptual metaphor?" Vyvyan Evans and Stéphanie Pourcel (eds.) *New directions in cognitive linguistics.* Amsterdam: John Benjamins: 127-145.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. L. Erlbaum Associates, Hillsdale, NJ.

Darwin, Charles. 1859. *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. John Murray, London

Darwin, Charles. 1871. *The descent of man, and selection in relation to sex*. John Murray, London

Engelbart, Douglas Carl. 1995. "Toward augmenting the human intellect and boosting our collective IQ." *Communications of the ACM*, *38*(8), 30-33.

English WK, Engelbart DC, and Bonnie Huddart. 1965. "Computer-Aided Display Control." *Final Report, Contract NASl-3988, SRI Project*, *5061*(1).

Gause Georgij F. 1934. *The struggle for existence*, Baltimore, Williams and Wilkins.

Girard, Gabriel. 1718. *La justesse de la langue françoise ou les différentes significations des mots qui passent pour synonimes*, Paris, Laurent d'Houry.

Goldberg, Adele E. 2019. *Explain Me This: Creativity, Competition, and the Partial Productivity of Constructions*. Princeton University Press.

Grice, H. Paul. 1975. "Logic and conversation." In Cole, P., and Morgan, J.(Eds.). *Syntax & Semantics 3*: 41-58.

Haeckel, Ernst. 1866. *Generelle Morphologie der Organismen. Allgemeine Grundzüge der organischen Formen-Wissenschaft, mechanisch begründet durch die von Charles Darwin reformirte Descendenz-Theorie*. Band 2. Berlin: Reimer.

Hagoort, Peter, Giosuè Baggio, and Roel M. Willems. 2009. "Semantic unification." *The cognitive neurosciences, 4th ed.*. MIT press, 819-836.

Hughes, Geoffrey. 1991. *Swearing: A Social History of Foul Language, Oath and Profanity in English*. Cambridge, MA: Blackwell.

Kant, Immanuel, 1781 [1881], *Critique of Pure Reason*. Translated by Friedrich Max Müller as *Immanuel Kant's Critique of Pure Reason: In Commemoration of the Centenary of Its First Publication*. Macmillan and Company, 1881.

Kendall, Maurice G. 1945. "The treatment of ties in ranking problems." *Biometrika* 33.3: 239-251.

Kerremans, Daphné, and Jelena Prokić. 2018. "Mining the Web for New Words: Semi-Automatic Neologism Identification with the NeoCrawler." *Anglia* 136.2: 239-268.

Lewis, David. 1969. *Convention: a philosophical study*. Cambridge, MA: Harvard University Press.

Lai, Randy. 2018. "arrangements: Fast Generators and Iterators for Permutations, Combinations and Partitions." R package version 1.1.5. https://CRAN.R-project.org/package=arrangements

Lee E.C., Whitehead AL, Jacques RM, et al. 2014. "The statistical interpretation of pilot trials: should significance thresholds be reconsidered?" *BMC Medical Research Methodology* 14: 41.

Mesoudi, Alex, Andrew Whiten, and Kevin N. Laland. 2006. "Towards a unified science of cultural evolution." *Behavioral and Brain Sciences* 29.4: 329-347.

Moseley, Tom. 2016. "The rise of the word Brexit". *BBC News, 25 December 2016*. https://www.bbc.com/news/uk-politics-37896977

Müller, Max. 1870. Darwinism tested by the Science of Language (translation 1869 of Schleicher 1863). Review. *Nature* 1, 256-259

Obermaier, Frederik, and Bastian Obermayer. 2017. "Wie die Paradise Papers entstanden. Antworten auf die wichtigsten Fragen rund um das neue Datenleck, das der SZ zugespielt wurde." *Süddeutsche Zeitung, Dienstag, 7.November 2017, Nr.255*, S.14.

Pashler, Harold, and Eric Jan Wagenmakers. 2012. "Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence?". *Perspectives on Psychological Science.* **7** (6): 528–530.

Peels, Rik. 2019. "Replicability and replication in the humanities." *Research integrity and peer review* 4.1: 2.

Pinker, Steven, and Paul Bloom. 1990. "Natural language and natural selection". *Behavioral and Brain Sciences.* 13 (4): 707–727. doi:10.1017/S0140525X00081061.

Progovac, Ljiljana. 2019. *A Critical Introduction to Language Evolution: Current Controversies and Future Prospects*. Springer Expert Briefs in Linguistics. Switzerland: Springer.

R Core Team. 2018. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. URL https://www.R-project.org/

Richards, Robert J. 2008. *The Tragic Sense of Life: Ernst Haeckel and the Struggle over Evolutionary Thought*. Chicago and London: University of Chicago Press.

Richerson, P. J. and Boyd, R. 2005. *Not by genes alone: How culture transformed human evolution*. Chicago and London: University of Chicago Press.

Schafer, J. L. and Graham, J. W. 2002. Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147– 177

Schleicher, August.1863. *Die Darwinsche Theorie und die Sprachwissenschaft – offenes Sendschreiben an Herrn Dr. Ernst Häckel, a. o. Professor der Zoologie und Director des zoologischen Museums an der Universität Jena*. Hermann Böhlau, Weimar.

Schleicher, August.1869. *Darwinism tested by the science of language. Translated by Dr. Alex V. W. Bikkers*. London: John Camden Hotten.

Schmid, Hans-Jörg. 2008. "New words in the mind: Concept-formation and entrenchment of neologisms." *Anglia-Zeitschrift für englische Philologie* 126.1: 1-36.

Smith, Richard. 1997. "Keeping the bad news from journalists." *British Medical Journal*, *314*: 81.

Sterrett, Susan G. 2002. "Darwin's Analogy Between Artificial and Natural Selection: How Does it Go?" *Studies in History and Philosophy of the Biological and Biomedical Sciences vol. 33 no. 1*: 151- 168.

Veale, Tony, and Cristina Butnariu. 2006. "Exploring Linguistic Creativity via Predictive Lexicology." In: *The Third Joint Workshop on Computational Creativity*, ECAI 2006.