Minimum cost and the emergence of the Zipf-Mandelbrot law

Paul Vogt[1,2]
[1]Language Evolution and Computation Research Unit, University of Edinburgh, U.K.
[2]Induction of Linguistic Knowledge / Computational Linguistics, Tilburg University, The Netherlands
paulv@ling.ed.ac.uk

Abstract

This paper illustrates how the Zipf-Mandelbrot law can emerge in language as a result of minimising the cost of categorising sensory images. The categorisation is based on the discrimination game in which sensory stimuli are categorised at different hierarchical layers of increasing density. The discrimination game is embedded in a variant of the language game model, called the selfish game, which in turn is embedded in the framework of iterated learning. The results indicate that a tendency to communicate in general terms, which is less costly, can contribute to the emergence of the Zipf-Mandelbrot law.

Introduction

One of the most sound universal tendencies observed in human languages is that when words are ranked according to their occurrence frequency in a descending order, the frequency f is inversely proportional to its rank k according to $f = Ck^{-B}$, where $B \approx 1$ and C is a constant, see Figure 1. This finding was discovered by G. K. Zipf (Zipf, 1949), and has since been called *Zipf's law*. Besides its observation in linguistics, Zipf's law has also been observed in economy, physics, biology, demography, social sciences etc. (Günther et al., 1996).

Zipf explained his finding in terms of least effort (Zipf, 1949). He assumed that speakers want to minimise articulatory effort, thus minimising the length of an utterance, which tends to promote ambiguity in language. On the other hand, hearers want to have optimal clarity to interpret the meaning of an utterance unambiguously with the least effort. Fulfilling the needs of both agents leads to a trade-off, which Zipf called the *principle of least effort* (Zipf, 1949). Although the observation of Zipf's law in real linguistic data is sound, it has only recently been shown empirically in an alife model that the principle of least effort indeed leads to a Zipfian distribution (Ferrer i Cancho and Solé, 2003).

In 1953, Mandelbrot derived a more general expression of Zipf's law, which explains small differences between Zipf's law and real linguistic data, notably for the first few ranks (Mandelbrot, 1953). According to the *Zipf-Mandelbrot law* the frequency f of a word is related to its rank k as $f = C(k+V)^{-B}$, where C, V and B are constants. When $V = 0$ and $B \approx 1$, this expression equals Zipf's law. Figure 1 shows both laws plotted on the usual log-log scale. It appears that Mandelbrot's equation fits linguistic data better than Zipf's equation (Mandelbrot, 1953). The derivation of Mandelbrot's formula was based on minimising the articulatory cost in terms of word length (Mandelbrot, 1953).[1]

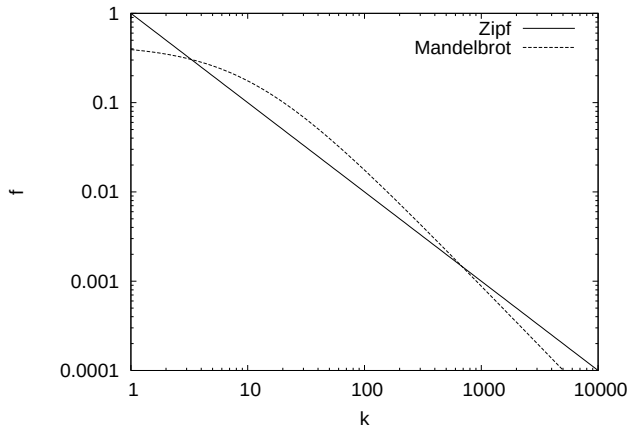

Figure 1: This plot shows both Zipf's law and the Zipf-Mandelbrot law. Mandelbrot's formula was drawn with parameters $C = 10$, $V = 10$ and $B = 1.35$ (these parameters are illustrative and not justified). The graph shows the frequency distribution f as a function of the rank k.

In addition to Zipf's least effort and Mandelbrot's minimum cost explanations, many other explanations of Zipf's law have been proposed that, too, focus on the relation between word length and its frequency. For example, it has been shown that randomly generated texts exhibit Zipf's law (Li, 1992), and so does the frequency distribution with which monkeys press the keys of a typewriter (Miller, 1957). This was explained by noting that generating shorter se-

[1]I sometimes refer to Zipf's law when I refer to the phenomenon that ranked words occur with are frequency distribution described by a hyperbola, irrespective whether expressed by Zipf's or Mandelbrot's equation. Where relevant, the distinction will be made.

quences is more statistically likely than generating long ones (Li, 1992). Explanations that do not explain the emergence of Zipf's law in terms of articulatory effort focus on frequency effects. More frequently used words are more likely to be selected in communication (Günther et al., 1996). This effect has been shown in a simulation where speakers select words based on occurrences in the preceding discourse (Tullo and Hurford, 2003).

This paper investigates the emergence of a Zipfian distribution in language as the result of a minimum cost principle (Mandelbrot, 1953), based on Steels' language game model (Steels et al., 2002). However, in contrast to Mandelbrot's derivation, the cost will not be minimised by optimising the word length, but rather by trying to minimise computational costs at the cognitive level of categorisation. As common in language game models, categorisation will be done using *discrimination games* (Steels, 1996). A recent discovery (Vogt, 2004) revealed that Zipfian distributions of word frequencies had emerged in robotic studies based on the language games, such as reported in (Vogt, 2000). This paper studies the hypothesis, suggested in (Vogt, 2004), that the emergence of Zipf's law may be explained by a tendency to use general categories in communication as a principle of minimum cost.

The next section outlines the model with which the study was done. Then the results are presented, which are discussed in the subsequent section. The final section provides conclusions.

The model

The study was done using the simulation toolkit THSim (Vogt, 2003b), which mimics aspects of the Talking Heads experiment (Steels et al., 2002).[2] THSim implements a number of different language games that can be incorporated by a population of agents. In the current study, the population plays *selfish games* – independently developed by Smith and Vogt (Smith, 2003; Vogt, 2000) – where hearers guess the reference from utterances produced by speakers, and learning is achieved by cross-situational statistical learning (Vogt and Smith, 2004). By engaging in selfish games, agents develop a repertoire of categories, which form the meanings of word-forms the agents develop as part of the selfish game.

The exact details of the selfish games are irrelevant for the purposes of this paper, similar results have been observed with – on language games based – guessing games (unpublished) and observational games (Vogt, 2004).[3] Figure 2 illustrates the working of the selfish games. In a selfish game, two agents – a speaker and a hearer – are selected from the population. Both agents look at a context

[2]The THSim toolkit containing the code of the present study is available at http://www.ling.ed.ac.uk/~paulv/thsim.html.

[3]Consult, e.g., (Vogt, 2000; Vogt, 2003b) for a description of these language game models.

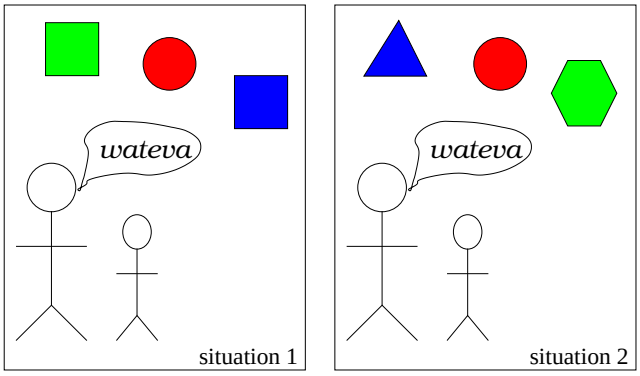

Figure 2: Two situations of selfish games illustrate the working of cross-situational learning. When a learner first hears the word "wateva" in a context with squares and a circle (situation 1), and later in a situation with a circle, a triangle and a polygon (situation 2), then she can induce – based on co-occurrences – the knowledge that "wateva" refers to the circle.

(or situation) that contains a number of coloured geometrical shapes. The speaker selects one shape as the topic, tries to categorise this topic and produces an utterance to convey its reference. If the speaker has no way to express a particular category (or meaning), she invents a new word-form. When the hearer receives an utterance, she tries to interpret the utterance by guessing which shape the speaker intends to convey. The hearer categorises all shapes and searches the association between the utterance and category that best matches the utterance in the given situation. This selection is based on maximising the probability $P(m|w)$ that given an utterance w, it means meaning m, provided the reference of the meaning is in the context. These probabilities are estimated according to word-meaning occurrences in previous situations. This learning mechanism has been called *cross-situational statistical learning* (Vogt and Smith, 2004) and works on the same principle as the cross-situational learning model introduced by Siskind (Siskind, 1996).

The selfish games are embedded in a cultural evolution where the language originates and is transmitted from one generation to the next culturally, i.e., the agents' morphologies remain the same throughout the course of evolution. At the start of each agent's lifetime, her linguistic knowledge is non-existent; this develops ontogenetically. The evolution is modelled using the iterated learning model (Kirby, 2002), which implements the population dynamics through iterating large sequences of selfish games played by the population. In each iteration, where a given number of selfish games are played, the population contains adult and learner agents. The adults are assumed to have mastered the language, which the learners learn by acting as hearers in selfish games with an adult as speaker. At the end of each iteration, the adults 'die' and are replaced by the learners, and new

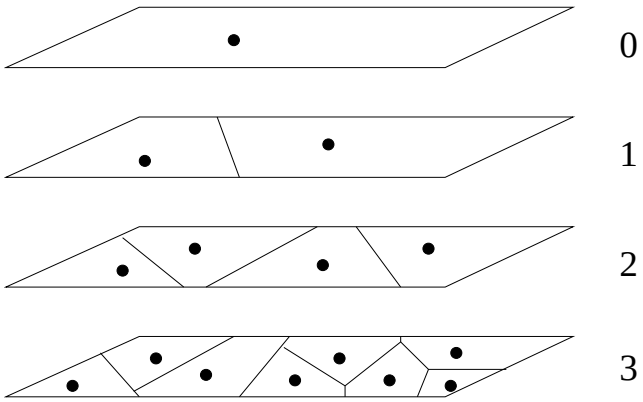

Figure 3: The hierarchical layering of categories using a prototype representation. Each layer l accepts up to n^l categories, which form Voronoi segments in the n-dimensional conceptual space. In the example here $n = 2$.

learners enter the population. This continues, in principle, indefinitely. The remainder of this section explains the categorisation process of the discrimination game in more detail.

The discrimination game

The agents categorise and form meanings using the discrimination game model proposed by Luc Steels (Steels, 1996). The aim of an agent playing a discrimination game is to categorise an object (the *topic*) such that it distinguishes the topic from all other objects in the context. In the original implementation (Steels, 1996), categories were represented by combinations of nodes in a binary tree. In the current implementation, categories are represented by prototypes that are points in an n-dimensional *conceptual space* (Gärdenfors, 2000), where the dimensions are *quality dimensions* which can be measured by feature detectors. For the present study, the conceptual space is 6-dimensional and is spanned by the R, G and B channels of the RGB colour space, a shape feature[4] and the x and y coordinates.

The distribution of prototypes leads to a segmentation of the space into Voronoi areas formed by the regions nearest to the prototypes; these Voronoi segments constitute the categories. A hierarchical layering of prototypes allows the emergence of a *taxonomic hierarchy* of categories (Rosch, 1978), see Fig. 3. In this taxonomy, categories on layers of low density are more general than those that are on layers of high density. The density $D(l)$ of layer l is given by $D(l) \leq n^l$, where n is the dimension of the conceptual space $S(l)$. The cost of categorisation is proportional to the time required to find a category that distinguishes the topic from the rest of the context. In the layered model, the agents can minimise categorisation cost by searching the different layers from the least dense layer to the more dense ones until a distinction can be made – i.e., the agents try to find the most general categories to be used in the communication act.

When an agent participates in the selfish game, she looks at the context C, which contains a number of objects ($|C| = 4$). The objects are selected randomly with a *uniform distribution* from a set of 10 different shapes, they are combined with an arbitrarily selected colour from a set of 11 colours and are placed at an arbitrarily selected point on the 2-dimensional display. For each object $o_i \in C$, the agent extracts a 6-dimensional feature vector $\mathbf{f}_i$ describing the objects in terms of its quality dimensions as mentioned above.

The speaker of the selfish game now selects a topic $o_t \in C$ and plays a discrimination game starting at layer $l = 1$ and continues at the next layer until $l = l_{max}$ or until the speaker found a *distinctive category*, see below. (l_{max} is the final layer that has at least one category for increasing values of l.) Given this category, the speaker tries to produce an utterance by searching its lexicon for a matching word-meaning association. If no such association is found, the speaker continues at the next layer until $l = l_{max}$. When the hearer receives an utterance, she plays a discrimination game for each object $o_i \in C$ starting at layer $l = 1$ and tries to interpret the utterance at layer l. This continues until the hearer interprets the utterance or until $l = l_{max}$. Note that an utterance is interpreted when the hearer found an association in its vocabulary that unambiguously identifies the utterance with a meaning (read *distinctive category*) that is consistent with the result of the discrimination games at layer l. This does not necessarily means that the interpretation is correct, which is only the case if the identified object $o = o_t$.

The discrimination game works as follows:

1. The feature vectors $\mathbf{f}_i$ of all objects $o_i \in C$ are categorised using the 1-nearest neighbourhood search (Cover and Hart, 1967) applied to the conceptual space at layer l. This results for each vector $\mathbf{f}_i$ in a category c_i, represented by the prototype $\mathbf{c}_i$ nearest to $\mathbf{f}_i$.
2. The agent then verifies whether the topic's category c_t distinguishes the topic from all other objects in the context. This holds when there is a *distinctive category* (or meaning) $m_t = c_t$ for which $\neg\exists o_j \in C \backslash \{o_t\} : c_j = c_t$.
3. If there does not exists such a meaning, then add a new category c – for which the topic's feature vector $\mathbf{f}_t$ is taken as an exemplar (i.e., $\mathbf{c} = \mathbf{f}_t$) – to the first hierarchical layer l that has space (i.e., $D(l) < n^l$) and return with failure.
4. Otherwise the category's prototype $\mathbf{c}_t$ is moved toward $\mathbf{f}_t$ such that it becomes the centre of mass of the feature vectors it distinctively categorised and return the distinctive category m_t. Note that if two categories become closer than within a given threshold, they are merged.

Note that as the hearer has to guess the topic, she considers all objects $o_i \in C$ as a potential topic and therefore plays a discrimination game for each potential topic at the given layer. This yields a distinctive category set M, which con-

[4]The shape feature is a value proportional to the shape's area divided by the area of its smallest bounding box (Vogt, 2003b).

tains the meanings of those objects that have successfully been distinguished.

As the complexity of the search in layer l is of order $o(D(l))$, the computational cost increases exponentially at lower levels in the taxonomic hierarchy (i.e., increasing values of l).

Results

In order to test the hypothesis that minimising the categorisation costs can lead to the emergence of Zipf's law, an experiment with 2 different conditions was carried out.

Condition 1 No hierarchical layering of conceptual spaces was used (there was only one layer available to each agent – and the density of the conceptual space was limited by the agents' memory sizes, which limits were never reached).

Condition 2 The hierarchical layering as described above was present.

For both conditions 10 trials of the simulation were run for 10 iterations of 100,000 selfish games. The population in each each iteration contained a total of 10 agents (5 adults and 5 learners). All speakers were selected from the adult population, but in order to prevent the emergence of many different languages, only 90% of the hearers were learners, the others were adults, consult (Vogt, 2003a) for a discussion.

Figure 4 summarises the most important results of the experiment. The two top figures and the leftmost figure on the bottom row show the ranked frequency distributions of 9 randomly selected agents throughout their lifetime (from each generation 1 agent) plotted on a log-log scale. In addition, these figures show the approximated curve of the Zipf-Mandelbrot equation with parameter settings as specified shortly.

The top left graph shows the results of condition 1, i.e., the run without hierarchical layering of categories. Clearly, the frequency distribution does not reveal the Zipf-Mandelbrot law $f = C(k+V)^{-B}$ with $B \approx 1$. This plot can be approximated by Mandelbrot's equation with $V \approx 1400$ and $B \approx 4$, which means that relatively many high ranked words occur almost equiprobable (V is high), while the occurrence frequencies of the remaining words drop faster than the Zipf-Mandelbrot law with $B \approx 1$ would predict.[5]

The top right graph of Fig. 4 shows the frequency of occurrences of word-meaning associations emerging under condition 2. This plot shows a curve similar to the Zipf-Mandelbrot curve (Fig. 1), which after a first small period transfers in an almost straight line with a slope near -1 ($V \approx 3$ and $B \approx 1.2$), which is typical for the Zipf-Mandelbrot law observed in natural languages. This figure shows the frequency distribution of *word-meaning* associations rather than the distribution of *word* occurrences. When the ranked frequency distribution of word occurrences is plotted against the rank k, an approximation of the Zipf-Mandelbrot law emerges with a value of $V \approx 80$ and $B \approx 5$, see Fig. 4 (bottom left).

In the condition 2 simulation, the categories of word-meanings that occupy conceptual spaces of higher density (i.e., higher values of l) have lower frequencies, which confirms the hypothesis that reducing the computational cost of categorisation can lead to the emergence of the Zipf-Mandelbrot law, see Fig. 4 (bottom right).

Not shown in the graphs are the communicative success of the two experiments. Communicative success measures the average number of successful selfish games over some window of time (a selfish game is successful if the hearer guessed the right topic). The simulation of condition 1 leads to an average communicative success of 69.3 ± 0.3 % over the final 10,000 games, whereas condition 2 yields an average of 51.2 ± 0.3 %. Note that the communicative success is averaged over the final 10,000 games of the simulation and averaged over all 10 trials.

Discussion

Minimising the computational costs of categorisation by trying to generalise the categorisation as much as possible does, indeed, lead to the emergence of the Zipf-Mandelbrot law (Mandelbrot, 1953). Moreover, Mandelbrot's equation matches the results better than Zipf's original equation. That the emergence of the Zipf-Mandelbrot law is caused by minimising the cost of categorisation through the hierarchical layering of conceptual spaces is prominently visible in the bottom right graph of Fig. 4, as the categories of lower frequency tend to occupy conceptual spaces of higher density.

The Mandelbrot-Zipf law emerges only for the ranked frequency distributions of the occurrence of *word-meaning associations*, rather than those of *words*, because as an artifact of the language game model, words tend to have multiple associations with different meanings at different layers. Nevertheless, these different meanings refer to the same objects in different situations. In human language, however, meanings at different hierarchical layers tend to have different words – e.g., animal, dog and spaniel. Assuming we can better model a one-to-one bias in word-meaning associations, see, e.g., (Smith, ming), thus specifying different taxonomies with more specific words, I decided to look at word-meanings as atomic elements rather than at words alone.

The agents (both as speaker and as hearer) do not optimise the effectiveness of their language use, but rather minimise the computational cost of categorisation. This aspect is a prominent reason why the communicative success of condi-

[5] The parameter values given are rough estimations; they are not obtained from statistical analysis, but the solid lines in the plots of Fig. 4 show the corresponding curve.

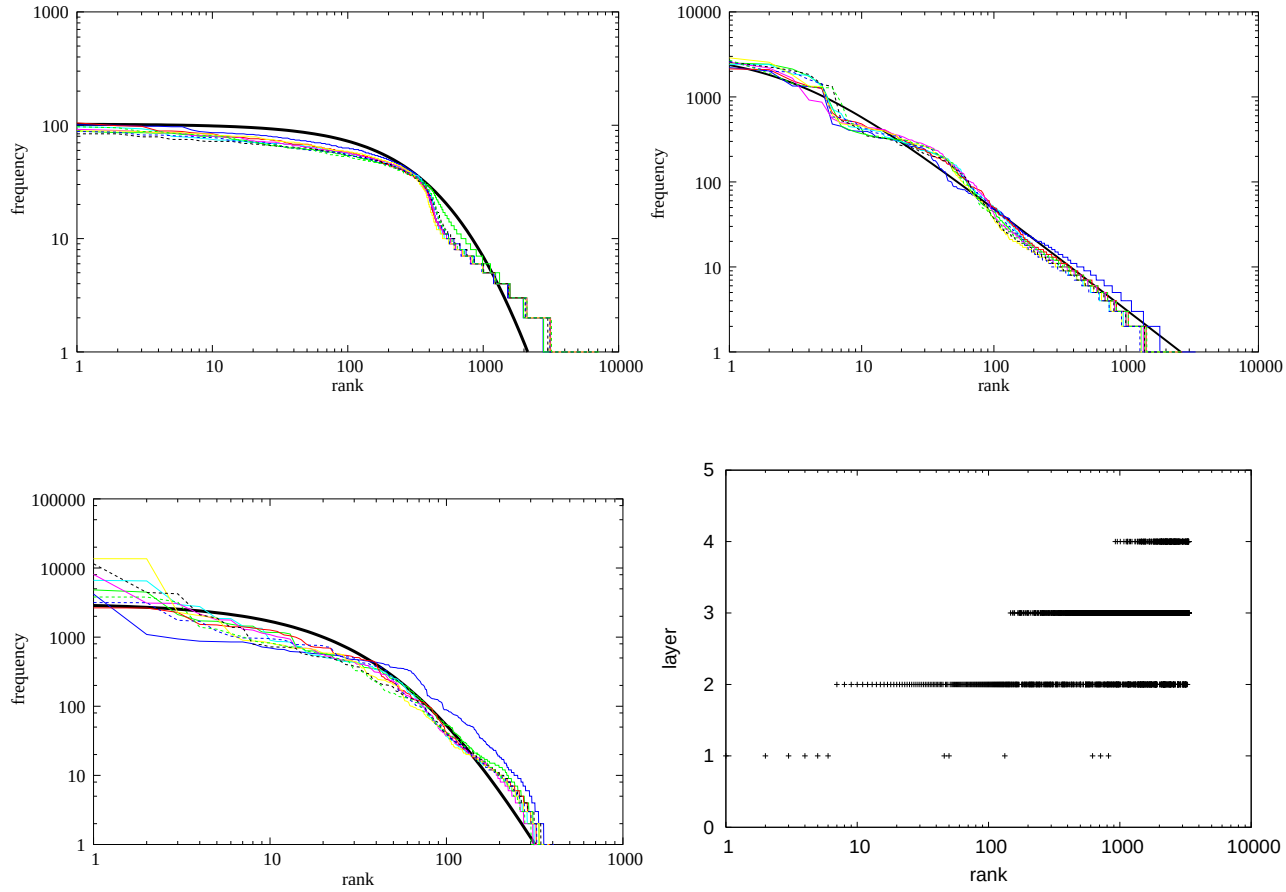

Figure 4: **Top left:** The frequency distribution of the occurrences of *word-meaning associations* as a function of the rank in condition 1. **Top right:** The frequency distribution of the occurrences of *word-meaning associations* in condition 2. **Bottom left:** The frequency distribution of *word* occurrences in condition 2. **Bottom right:** The hierarchical layers l occupied by the categories of rank k relating to one of the agents in condition 2 from the previous two graphs. The first three graphs additionally shows the Mandelbrot equation imposed with parameter settings as given in the text.

tion 2 stays well behind the success of condition 1, where the agents do search the entire lexicon for associations, thus including those that allow more effective communication. Hence, minimising the cost of categorisation while not optimising communicative success causes a trade-off between effort and understandability, which was the basis of Zipf's *principle of least effort* (Zipf, 1949). It is likely that Zipf's law will not emerge when the agents optimise for effectiveness, which – of course – is costly, when a hierarchical taxonomy of categories is maintained. However, the current level of success is too low to be acceptable. Future work should investigate how the success of the communication can increase without increasing the cost too much.

Conclusions

This paper shows that the Zipf-Mandelbrot law can emerge as a result of minimising the cost of categorising sensory stimuli. The emergence is striking, because no aspect of a Zipfian distribution was put in the model: not in the distribution of objects the agents categorise, nor in the distribution of categories in the hierarchically layered conceptual spaces.

To investigate the validity of the general hypothesis that minimising the cost of categorisation does lead to the emergence of Zipf's law in human language, it would be interesting to investigate to what extent the shorter words used in real languages are indeed the more general terms. If that is the case, it should also be investigated to what extent the more general categories are indeed cognitively less costly to categorise, such as appears to be the case for basic level categories as opposed to their superordinate and subordinate categories (Rosch, 1978).

It is important to stress that I do not claim that minimising cost of categorisation is *the* mechanism for the observation of the Zipf-Mandelbrot law in natural languages; there

are too many possible explanations around to make such a hard claim (Günther et al., 1996). Furthermore, the most frequently used words are function words, such as 'the' and 'a', which carry no meaning in the sense used here (although one might argue that these words have very general meanings, or at least are applicable in a very general way).

However, humans undoubtedly try to minimise the cognitive effort to categorise sensorimotor events. It is therefore plausible that the tendency to use generalised categories is a bias that – *in addition to other biases*, such as reducing articulatory effort (Mandelbrot, 1953; Zipf, 1949) and other frequency related approaches (Günther et al., 1996; Tullo and Hurford, 2003) – yields the emergence of the Zipf-Mandelbrot law.

Acknowledgements

This work was supported by a VENI grant provided by the Netherlands Organisation for Scientific Research (NWO). Many thanks go to Martin Reynaert and Andrew Smith and the four anonymous reviewers for their useful suggestions and comments on this work.

References

Cover, T. M. and Hart, P. E. (1967). Nearest neighbour pattern classification. *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, 13:21–27.

Ferrer i Cancho, R. and Solé, R. V. (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences*, 100(3):788–791.

Gärdenfors, P. (2000). *Conceptual Spaces*. Bradford Books, MIT Press.

Günther, R., Levitin, L., Schapiro, B., and Wagner, P. (1996). Zipf's law and the effect of ranking on probability distributions. *International Journal of Theoretical Physics*, 35(2):395–417.

Kirby, S. (2002). Learning, bottlenecks and the evolution of recursive syntax. In Briscoe, T., editor, *Linguistic Evolution through Language Acquisition: Formal and Computational Models*. Cambridge University Press.

Li, W. (1992). Random texts exhibit Zipf's law like word frequency distributions. *IEEE Transactions on Information Theory*, 38(6):1842–1845.

Mandelbrot, B. B. (1953). An information theory of the statistical structure of language. In Jackson, W., editor, *Communication Theory*, pages 503–512, New York. Academic Press.

Miller, G. A. (1957). Some effects of intermittent silence. *American Journal of Psychology*, 70:311–314.

Rosch, E. (1978). Principles of categorization. In Rosch, E. and Lloyd, B. B., editors, *Cognition and Categorization*. Lawrence Erlbaum Ass.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, 61:39–91.

Smith, A. D. M. (2003). Intelligent meaning creation in a clumpy world helps communication. *Artificial Life*, 9(2):559–574.

Smith, K. (forthcoming). The evolution of vocabulary. *Journal of Theoretical Biology*.

Steels, L. (1996). Perceptually grounded meaning creation. In Tokoro, M., editor, *Proceedings of the International Conference on Multi-Agent Systems*, Menlo Park Ca. AAAI Press.

Steels, L., Kaplan, F., McIntyre, A., and Van Looveren, J. (2002). Crucial factors in the origins of word-meaning. In Wray, A., editor, *The Transition to Language*, Oxford, UK. Oxford University Press.

Tullo, C. and Hurford, J. R. (2003). Modelling Zipfian distributions in language. In Kirby, S., editor, *Language Evolution and Computation, Proceedings of the workshop at ESSLLI*.

Vogt, P. (2000). Bootstrapping grounded symbols by minimal autonomous robots. *Evolution of Communication*, 4(1):89–118.

Vogt, P. (2003a). Grounded lexicon formation without explicit meaning transfer: who's talking to who? In Banzhaf, W., Christaller, T., Dittrich, P., Kim, J. T., and Ziegler, J., editors, *Advances in Artificial Life - Proceedings of the 7th European Conference on Artificial Life (ECAL)*. Springer Verlag Berlin, Heidelberg.

Vogt, P. (2003b). THSim v3.2: The Talking Heads simulation tool. In Banzhaf, W., Christaller, T., Dittrich, P., Kim, J. T., and Ziegler, J., editors, *Advances in Artificial Life - Proceedings of the 7th European Conference on Artificial Life (ECAL)*. Springer Verlag Berlin, Heidelberg.

Vogt, P. (2004). Generalisation as a bias toward the emergence of Zipf's law. In *Proceedings of Evolang 5*.

Vogt, P. and Smith, A. D. M. (2004). Quantifying lexicon acquisition under uncertainty. In Lenaerts, T., Nowe, A., and Steenhout, K., editors, *Proceedings of Benelearn 2004*.

Zipf, G. K. (1949). *Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley, Cambridge, MA.