Measuring the Productivity of Determinerless PPs

Florian Dömges, Tibor Kiss, Antje Müller, Claudia Roch
Sprachwissenschaftliches Institut
Ruhr-Universität Bochum

florian.doemges@rub.de
tibor@linguistics.rub.de
antje.mueller@rub.de
claudia.roch@rub.de

Abstract

We determine the productivity of determinerless PPs in German quantitatively, restricting ourselves to the preposition *unter*. The study is based on two German newspaper corpora, comprising some 210 million words. The problematic construction, i.e. *unter* followed by a determinerless singular noun occurs some 16.000 times in the corpus. To clarify the empirical productivity of the construction, we apply a productivity measure developed by Baayen (2001) to the syntactic domain by making use of statistical models suggested in Evert (2004). We compare two different models and suggest a gradient descent search for parameter estimation. Our results show that the combination of *unter*+noun must in fact be characterized as productive, and hence that a syntactic treatment is required.

1 Introduction

The combination of a preposition with a singular count noun, illustrated in (1) with the preposition *unter*, is a frequent construction in written and spoken German. From a theoretical perspective, constructions like (1) are interesting since they seem to violate the near universal rule that determiners should accompany singular count nouns if the language in question shows determiners at all (cf. Himmelmann (1998)).

unter Vorbehalt (with reservation),
unter Androhung (on pain),
unter Lizenz (under licence),
unter Vorwand (pretending) (1)

Baldwin et al. (2006) follow a tradition of English grammar and call constructions like (1) determinerless PPs (D-PP), defined as PPs whose NP-complement consists of a singular count noun without an accompanying determiner (as e.g. English *by bus*, *in mind*). It has been claimed that D-PPs are mostly idiomatic and not productive. Hence, computational grammars often include D-PPs only as stock phrases or listed multiword expressions and do not offer a grammatical treatment. However, both claims have to be doubted seriously. Kiss (2006, 2007) shows that the class of D-PPs does not contain more idiomatic phrases than a typical phrasal category should and also argues against a 'light P hypothesis' which allows a pseudo-compositional treatment of D-PPs by ignoring the semantics of the preposition altogether. Trawinski (2003), Baldwin et al. (2006), as well as Trawinski et al. (2006) offer grammatical treatments of D-PPs, or at least of some subclasses of D-PPs. Interestingly, (Baldwin et al. (2006), 175f.) take the productivity of a subclass of D-PPs for granted and propose a lexical entry for prepositions which select determinerless N's as their complement. While we are sympathetic to a syntactic treatment of D-PPs in a computational grammar, we think that the productivity of such constructions must be considered more closely. The analysis of Baldwin et al. (2006) allows the unlimited combination of prepositions meeting their lexical specification with a determinerless N projection. This

Proceedings of the 4th ACL-SIGSEM Workshop on Prepositions, pages 31–37,
Prague, Czech Republic, June 2007.

assumption is not in line with speaker's intuitions with regard to producing or judging such constructions. As has been pointed out by Kiss (2006, 2007), speakers of German can neither freely produce sequences consisting of *unter* and determinerless N projections (typically a noun) nor can they judge such constructions in isolation. In addition, not even very similar nouns can be interchanged in a D-PP, as can be witnessed by comparing near-synonyms *Voraussetzung* and *Prämisse* which both translate as prerequisite, or as provided in the examples in (2).

The examples in (2) illustrate that *Voraussetzung* cannot be replaced by *Prämisse* in a D-PP (2a, b), while it can be replaced as a head noun in a full PP (2c, d). While the contrast in (2) casts doubt on a productive analysis on the basis of the speakers knowledge of language, the present paper will show that *unter*+noun has to be classified as productive from an empirical perspective.

(2)
a. Auch Philippe Egli besteht auf einer eigenen Handschrift - *unter Voraussetzung* des Einverständnisses des Ensembles.
b. * Auch Philippe Egli besteht auf einer eigenen Handschrift - *unter Prämisse* des Einverständnisses des Ensembles.
c. Auch Philippe Egli besteht auf einer eigenen Handschrift - *unter der Voraussetzung* des Einverständnisses des Ensembles.
d. Auch Philippe Egli besteht auf einer eigenen Handschrift - *unter der Prämisse* des Einverständnisses des Ensembles.

"Philippe Egli insists on his individual way of dealing with the issue, provided the ensemble agrees."

Our investigation is based of a corpus analysis of D-PPs, consisting of the preposition *unter* and a following noun, and employs a quantitative measure of productivity, first developed by Harald Baayen to analyze morphological productivity. The preliminary conclusion to be drawn from this result will be that empirical and intuitive productivity of *unter*+noun sequences do not match.

In applying Baayen's productivity measure to syntactic sequences, however, we are faced with a serious problem. Baayen's productivity measure $P(N)$ is based on the expectation of the hapax legomena – $E[V_1]$ – occurring in a vocabulary of size N, i.e. $P(N) = \frac{E[V_1]}{N}$.

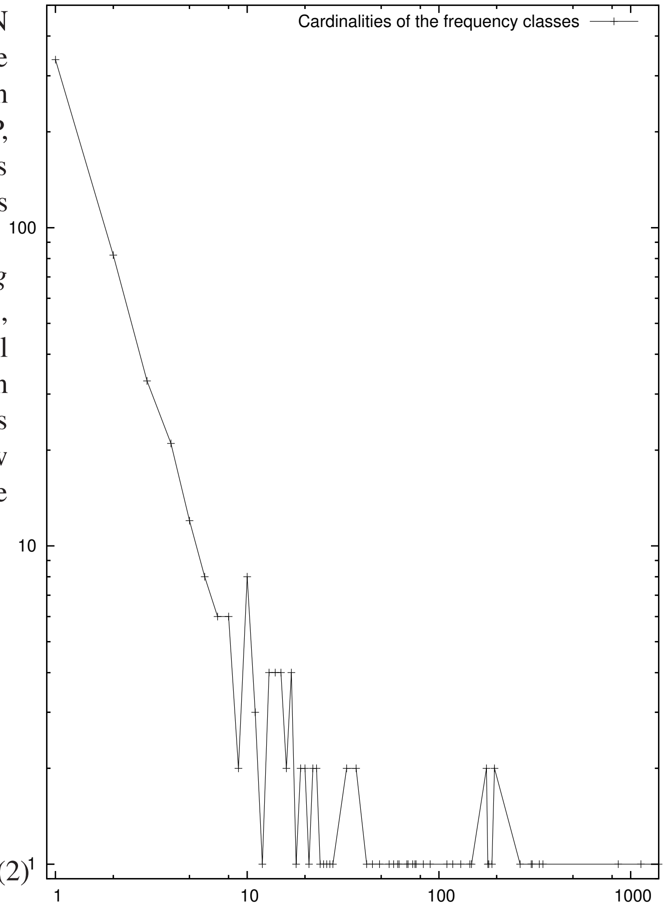

Figure 1: Cardinalities of the frequency classes. The frequency of each type was counted, then the types were grouped into classes of equal frequency. The number of types in each class was counted. The frequency values m are assigned to the x-axis, the size of the class V_m to the y-axis. Both are scaled logarithmically.

Since we cannot derive the expectation of the hapax legomena directly from the corpus, we have to approximate it by use of regression models. To simplify matters somewhat, Baayen's models can only be applied to unigrams, while we have to consider bigrams – the preposition and the adjacent noun. To circumvent this problem, Kiss (2006,2007) calculated $P(N)$ on the basis of the empirical distribution of V_1 as N gets larger. Evert (2004) offers regression models to determine $E[V_1]$ for n-grams and suggests two different models, the Zipf-Mandelbrot model (ZM) and the finite Zipf-Mandelbrot model (fZM). The difference between these two models is that fZM assumes a finite vocabulary. In the present paper, we apply Evert's models to sequences of *unter*+noun. We differ from Evert's proposal in estimating the free parameter α in both models on the basis of the gradient descent algorithm. Contrary to Evert's assumptions, we will show that the results of the ZM model are much closer to the empirical observations than the results of the fZM model.

The paper is structured as follows. Section 2 describes the empirical basis of the experiment, a corpus study of *unter*$+textnoun_{sg}$ sequences. Section 3 introduces the models suggested by Evert (2004). Section 3.1 introduces the models, section 3.2 shows how the free parameter is estimated by making use of the gradient descent algorithm. The results are compared in section 3.3.

2 Corpus Study

The present study is based on two German corpora, with a total of 213 million words: the NZZ-corpus 1995-1998 (Neue Zürcher Zeitung) and the FRR-corpus 1997-1999 (Frankfurter Rundschau). Making use of the orthographic convention that nouns are capitalized in German, we have automatically extracted 12.993 types, amouting to some 71.000 tokens of *unter* and a following noun. From these 12.993 types, we have removed all candidates where the noun is a proper noun, or realized as a plural, or as member of a support verb construction. Also, we have excluded typical stock phrases and all mass nouns. The extraction process was done both manually (proper nouns, mass nouns, support verb constructions) and automatically (plurals, mass nouns).

As a result of the extraction process, a total number of 1.103 types remained, amounting to 16.444 tokens. The frequency of every type was determined and types with the same frequency were grouped into classes. 65 equivalence classes were established according to their frequency m (cf. Figure 1). The number of elements in every class was counted and the various count results were associated with the variables $V_m = V_1, V_2, \ldots, V_{2134}$.

3 LNRE Model Regression

Baayen (2001) uses the term LNRE models (*large number of rare events*) to describe a class of models that allow the determination of the expectation with a small set of parameters. Evert (2004) proposes two LNRE models with are based on Zipf's Law (Zipf(1949), Li (1992)) to identify the expectations $E[V_1], \ldots, E[V_{max}]$. Both models are based on the Zipf-Mandelbrot law.

Zipf's Law (Zipf(1949), Li (1992)) posits that the frequency of the r-most frequent type is proportional to $\frac{1}{r}$. The distribution of random texts displays a strong similarity to the results expected according to Zipf's Law (cp. Li (1992)). Mandelbrot (1962) et al. explain this phenomenon by Zipf's *Principle of Least Effort*.

Rouault (1978) shows that the probability of types with a low frequency asymptotically behaves as posited by the Zipf-Mandelbrot Law

$$\pi_i = \frac{C}{(i+b)^a}$$

with $a > 1$ and $b > 0$.

The models are introduced in section 3.1. Both require a parameter α, whose value was determined by employing a gradient descent algorithm implemented in Perl. The optimal value for the free parameter was found by constructing an error function to minimise α. The calculation was carried out for both models, but better results are produced if the assumption is given up that the vocabulary is finite.

3.1 Finite and general Zipf-Mandelbrot models

Evert (2004) proposes the finite Zipf-Mandelbrot model (fZM) and the general Zipf-Mandelbrot model (ZM) for modelling the expectations of the frequency classes V_m, i.e. $E[V_1], \ldots, E[V_{max}]$ and the expected vocabulary size, i.e. the expectation of the different types $E[V]$. The two models make different assumptions about the probability distributions of the frequency classes. The fZM assumes that there is a minimal probability A – defined as $\exists A : \forall i : A \leq \pi_i$. This amounts to the assumption that the vocabulary size itself is finite. Hence, it can be expected according to the fZM model that the set of observed types does not increase once $N \approx \frac{1}{A}$ is reached. In the general ZM model, there is no such minimal probability.

Assuming a fZM model, Evert (2004) proposes the following results to estimate the expectation of the frequency classes $E[V_m]$ and the expected vocabulary size $E[V]$. In the following equations, B stands for the maximum probability, defined as $\forall i : B \geq \pi_i$.

$$E[V_m] = \frac{1-\alpha}{(B^{1-\alpha} - A^{1-\alpha}) \cdot m!} \cdot$$

$$N^{\alpha} \cdot \Gamma(m-\alpha, N \cdot A) \tag{3}$$

$$E[V] = \frac{1-\alpha}{(B^{1-\alpha} - A^{1-\alpha})} \cdot N^{\alpha} \cdot \frac{\Gamma(1-\alpha, N \cdot A)}{\alpha} +$$

$$\frac{1-\alpha}{(B^{1-\alpha} - A^{1-\alpha}) \cdot \alpha \cdot A^{\alpha}} \cdot (1 - e^{-N \cdot A}) \tag{4}$$

As can be witnessed from the formulae given, N, A, and B are already known or directly derivable from our observations, leaving us with the determination of the free parameter α.

Using the general Zipf-Mandelbrot model, we end with the following estimations, again suggested by Evert (2004):

$$E[V_m] = \frac{1-\alpha}{B^{1-\alpha} \cdot m!} \cdot N^{\alpha} \cdot \Gamma(m-\alpha) \tag{5}$$

$$E[V] = \frac{1-\alpha}{B^{1-\alpha}} \cdot N^{\alpha} \cdot \frac{\Gamma(1-\alpha)}{\alpha} \tag{6}$$

As there is no minimal probability, we are left with the maximal probability B, the token size N, and again a free parameter α.

3.2 Parameter estimation through gradient descent

Since the expectation of the frequency classes in (3) and (5) depend on the free parameter α, this parameter must be estimated in a way that minimises the deviation of expected and observed values. We measure the deviation with a function that takes into account all observed frequencies and their expected values. A function satisfying these criteria can be found by treating observed frequency classes and expectations as real-valued vectors in a vector space.

$$\mathbf{O}^T = (V, V_1, V_2, \ldots, V_{2134}) \in \mathbb{R}^{66} \tag{7}$$

$$\mathbf{E}^T(\alpha) =$$

$$(E(V)(\alpha), E(V_1)(\alpha), \ldots, E(V_{2134})(\alpha)) \in \mathbb{R}^{66} \tag{8}$$

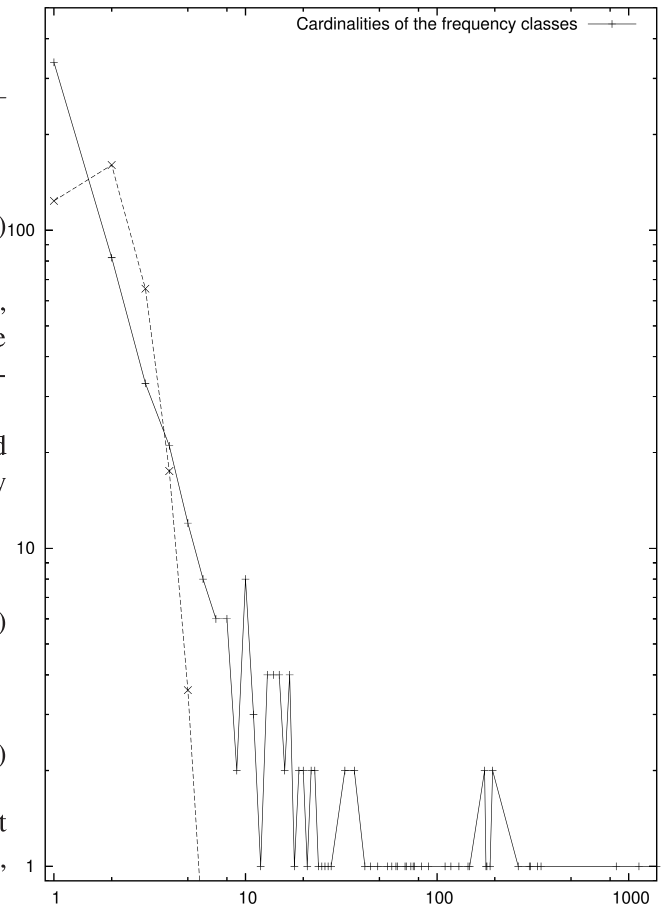

Figure 2: The application of the fZM LNRE Model combined with Rouault's estimation method leads to a strong deviation from the observed data. The observed data is depicted as a solid line, the data from the model as a dotted line. The frequency values m are assigned to the x-axis, the size of the class V_m respectively the expected size $E(V_m)$ to the y-axis. Both are scaled logarithmically.

A natural choice for a measure of error is the quadratic norm of the difference vector between observation and expectation. As we have no information about the relationship between different frequencies we assume that the covariance matrix is the unit matrix.

These thoughts result in the following error function:

$$g(\alpha) = (E(V)(\alpha) - V)^2 +$$

$$\sum_{m=1,\ldots,2134} (E(V_m)(\alpha) - V_m)^2 \quad (9)$$

The minimal α is equal to the root of the derivative of the error function with respect to α. The derivative of the error function is:

$$\frac{\partial g}{\partial \alpha} = 2\frac{\partial E(V)}{\partial \alpha}(E(V)(\alpha) - V) +$$

$$2 \sum_{m=1,\ldots,2134} \frac{\partial E(V_m)}{\partial \alpha}(E(V_m)(\alpha) - V_m) \quad (10)$$

One way to find the minimum $\alpha^* = argmin_\alpha\, g(\alpha)$ would be to derive the expected values with respect to α and solve $g'(\alpha^*) = 0$ for α. As there is no way known to the authors to accomplish this in a symbolic way, the use of a numeric method to calculate α^* is advised.

We chose to find α^* by employing a gradient descent method and approximating $\frac{\partial g}{\partial \alpha}$ by evaluating $g(\alpha)$ in small steps $\epsilon_\alpha(i)$ and calculating $\frac{\Delta g(k)}{\epsilon_\alpha(k)} = \frac{g(\alpha_0 + \sum_{j=1}^{k} \epsilon_\alpha(j)) - g(\alpha_0 + \sum_{j=1}^{k-1} \epsilon_\alpha(j))}{\epsilon_\alpha(k)}$, where k is number of the iteration.

In the vicinity of a minimum $\frac{\partial g}{\partial \alpha}(\alpha)$ decreases until it vanishes at α^*.

After every iteration the new $\epsilon_\alpha(k)$ is chosen by taking under consideration the change of $\frac{\Delta g(k)}{\epsilon_\alpha(k)}$ and the sign of $\epsilon_\alpha(k-1)$. If $\frac{\Delta g(k)}{\epsilon_\alpha(k)}$ increased, the sign of $\epsilon_\alpha(k-1)$ is inverted: $\epsilon_\alpha(k) = -\epsilon_\alpha(k-1)$.

To prevent the algorithm from oscillating around the minimum the last two values $g(\alpha_0 + \sum_{j=1}^{k-2} \epsilon_\alpha(j))$ and $g(\alpha_0 + \sum_{j=1}^{k-1} \epsilon_\alpha(j))$ are saved.

When a step would result in returning to a previous value $g(\alpha_0 + \sum_{j=1}^{k-1} \epsilon_\alpha(j) + \epsilon_\alpha(k)) = g(\alpha_0 + \sum_{j=1}^{k-2} \epsilon_\alpha(j))$, the step size is multiplied by a constant $0 < \gamma \leq 1$: $\epsilon_\alpha(k) = \gamma\epsilon_\alpha(k-1)$. The algorithm is stopped when the absolute value of the step size drops under a predetermined threshold: $|\epsilon_\alpha(k)| < \epsilon_{threshold}$.

3.3 Results

Interestingly, α as determined by gradient descent on the basis of a fZM leads to a value of 0.666, which does not match well with our observations, as can be witnessed in Figure 2.

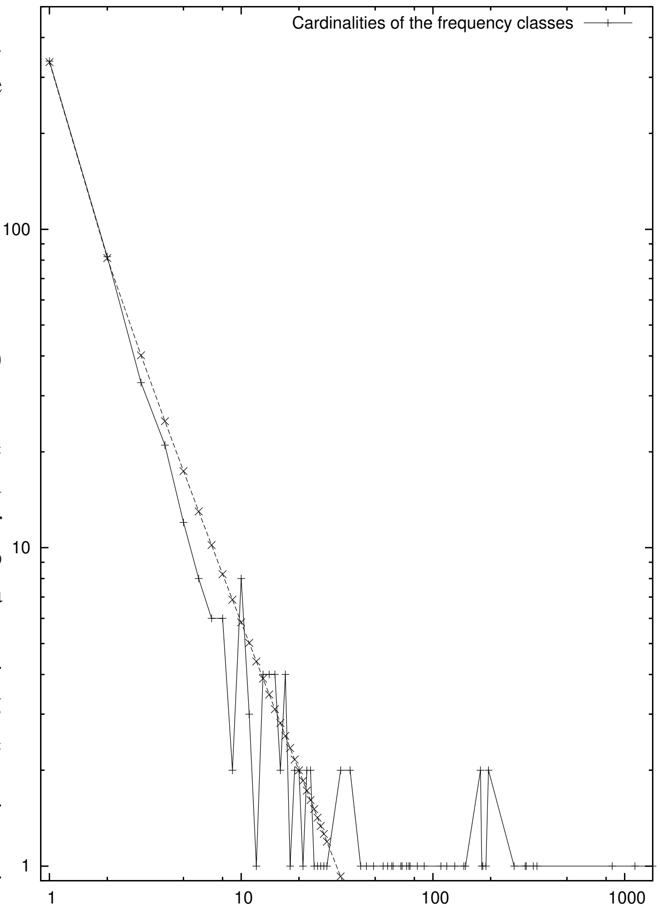

Figure 3: The ZM LNRE Model leads to a far better result with less deviation from the observation. The observed data is depicted as a solid line, the data from the model as a dotted line. The frequency values m are assigned to the x-axis, the size of the class V_m respectively the expected size $E(V_m)$ to the y-axis. Both are scaled logarithmically.

A gradient descent search on the basis of the ZM model delivered a value of $\alpha = 0.515$, a much better approximation (with a χ^2-Value of 4.514), as can be witnessed from Figure 3. The value thus reached also converges with the estimation procedure for α suggested by Rouault (1978), and taken up by Evert (2004), i.e. $\alpha = \frac{V_1}{V}$. Consequently, we assume a ZM model for estimating of expected frequencies.

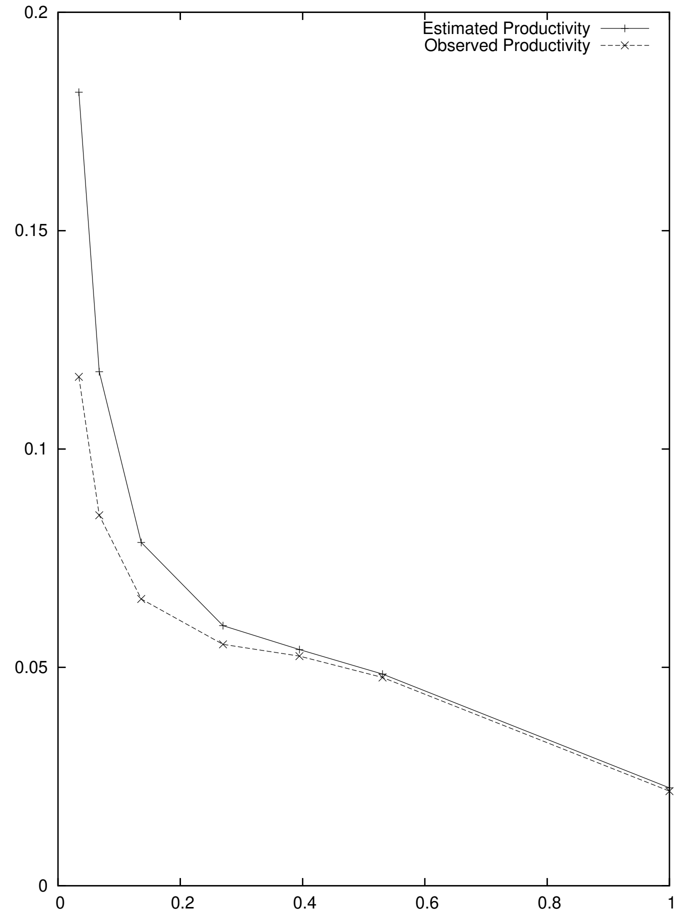

Figure 4: The parts of the corpus were appended to each other and after every step the productivity $P(N)$ was calculated directly from the data as well as from the fitted ZM model. The percentage of the corpus is assigned to the x-axis, the productivity $P(N)$ is assigned to the y-axis. The productivity values that were deduced directly from data are plotted as a dotted line, the productivity values from the ZM model are plotted as a solid line.

To chart the productivity of sequences of the form *unter*+noun, we have divided our corpus into six smaller parts and sampled V, N, and V_1 at these parts. The distribution of the observations thus gained can be found in Figure 4, together with the expectations derived from the ZM model. We observe that both distributions are strikingly similar and converge at the values for the full corpus.

N	V_1	$E[V_1]$	$P(N)$
542	74	96.66	0.182
1068	104	123.47	0.118
2151	169	166.41	0.079
4262	282	249.93	0.059
6222	384	332.19	0.054
8365	469	400.43	0.048
16444	746	748.81	0.022

Table 1: Overview of the observed and expected numbers of hapax legomena and the associated productivity value at different corpus sizes.

In a broader perspective, Figure 4 shows that the combination of *unter*+noun is a productive process, when its empirical distribution is considered. As was already pointed out in section 1, this finding is at odds with speaker's intuitions about combinations of *unter*+noun. Assuming that this result can be extended to other subclasses of D-PPs, we would suggest restricting lexical specifications for prepositions to subclasses of nouns, depending on the pertinent preposition. Future research will have to show whether such clear-cut subclasses can be identified by looking more closely at the empirical findings, other whether we are confronted with a continuum, which would require alternative rule types.

References

Harald Baayen. 2001. *Word Frequency Distributions*. Kluwer, Dordrecht.

Timothy Baldwin, John Beavers, Leonoor van der Beek, Francis Bond, Dan Flickinger, and Ivan A. Sag. 2006. *In Search of a Systematic Treatment of Determinerless PPs*. In Patrick Saint-Dizier, editor, *Syntax and Semantics of Prepositions*, pages 163–179. Springer.

Stefan Evert. 2004. *A Simple LNRE Model for Random Character Sequences*. In *Proceedings of the 7mes Journees Internationales d'Analyse Statistique des Donnees Textuelles*, pages 411–422.

Nikolaus Himmelmann. 1998. *Regularity in Irregularity: Article Use in Adpositional Phrases*. *Linguistic Typology*, 2:315–353.

Tibor Kiss. 2006. *Do we need a grammar of irregular sequences?* In Miriam Butt, editor, *Proceedings of KONVENS*, pages 64–70, Konstanz.

Tibor Kiss. 2007. *Produktivität und Idiomatizität von Präposition-Substantiv-Sequenzen. forthcoming in Zeitschrift für Sprachwissenschaft*.

W. Li. 1992. *Random texts exhibit zipf's-law-like word frequency distribution. IEEE Transactions on Information Theory*.

B. Mandelbrot. 1962. *On the theory of word frequencies and on related Markovian models of discourse*. American Mathematical Society.

A. Rouault. 1978. *Lois de Zipf et sources markoviennes. Annales de l'Institut H. Poincare*.

Beata Trawinski, Manfred Sailer, and Jan-Philipp Soehn. 2006. *Combinatorial Aspects of Collocational Prepositional Phrases*. In Patrick Saint-Dizier, editor, *Syntax and Semantics of Prepositions*, pages 181–196. Springer.

Beata Trawinski. 2003. *The Syntax of Complex Prepositions in German: An HPSG Approach.* In *Proceedings of GLIP*, volume 5, pages 155–166.

G. K. Zipf. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Campridge.