# A Stable Statistical Constant Specific for Human Language Texts

Felix Golcher
Humboldt-Universität zu Berlin
Unter den Linden 6
10099 Berlin, Germany
felix.golcher@hu-berlin.de

## Abstract

A novel character-level statistical measure is described which quantifies the level of repetitions in a text. It behaves remarkably uniformly for texts from all 20 tested languages. In contrast to most other text-statistical quantities, the proposed measure is computed from the text as a whole, not from a tokenised text reduced to a frequency list. For growing text sizes, it converges rapidly to a constant value. This text length independent behaviour is an uncommon feature for text-statistical constants. The described phenomenon of constant repetitiveness has so far not been observed in any non-natural language text.

## Keywords

Text statistics; lexical constant; language universal; suffix trees

## 1 Introduction

Since George Kingsley Zipf first published the famous empirical law since named after him [13], a lot of text statistical regularities have been proposed, usually in the form of a formula with some constants in it. (See [2] for an overview.)

However, recent publications have raised doubts as to whether these laws hold and whether these constants are constant. In [11], it is shown for all alleged lexicostatistic constants known at the time that they systematically depend on text size. Additionally, Evert and Baroni [5] demonstrate the low predictive power of many of the laws that were proposed to cope with such text length dependencies.

The *theoretical* significance of Zipf's Law and its relatives is limited by three factors: firstly, they merely make propositions about word frequency lists instead of full human texts. Secondly, they apply in a very similar fashion to randomly produced pseudo text [9, 8] and thus are not a specific property of language. Thirdly, they are not easily interpreted theoretically: it's unclear what the validity of Zipf's Law actually tells us about the system of natural language and its properties.

This paper introduces a new text statistical measure $V$ which quantifies the level of repetitiveness in a text. For natural language text, $V$ converges rapidly towards a fixed value, as the text size grows, and the convergence point is a good constant over texts from different languages. This was tested with 20 languages, from three distinct language families, written with three different classes of writing systems.

So far, this constant repetition rate has only been observed for natural language text. The possible establishing of this phenomenon of constant repetitiveness as a universal and exclusive feature of human text could have some impact on the theory of language: on the one hand, it would impose restrictions on every realistic language model, since such a model would have to reproduce this property in its output (see the discussion in Section 6). On the other hand, the phenomenon would bring up two new questions: if the level of repetitiveness is so amazingly constant, why is this so and what mechanism keeps it constant?

Section 2 gives the necessary conceptual background and defines $V$. Section 3 describes the experiments which survey $V$ for texts from different languages. The results are shown in Section 4. Section 5 reports an experiment which gives more insight into the nature of the investigated quantity. Section 6 discusses comparable known text statistical measures. Section 7 gives an outlook.

## 2 The measure $V$

### 2.1 Defining $V$

In the context of this work, a text is simply a string of symbols. I define the repetitiveness $V$ of a text $T$ as $k/t_0$, where $t_0$ is the length of $T$, and $k$ is the number of its substrings which occur with more than one continuation in $T$. In other words, $V$ is the number of *ended* repetitions divided by the text length.

Consider the example text[1] $T =$ `xabcdecdeabcbx`. There are 7 substrings with more than one continuation: `abc`, `bc`, `c`, `cde`, `de`, `e`, and `b`. The text length $t_0$ is 14, hence $V = 7/14 = 1/2$.

If there are no repetitions in the text, $V$ is obviously 0. If the text consists of the same character repeating – except for the last character – then $k = t_0 - 2$ and thus $V = (t_0 - 2)/t_0$, which approaches 1 as $t_0$ grows.

Quantifying the repetitiveness of a text by defining $V$ can be justified a priori by its conceptual simplicity and adequateness (it measures repetitions). Its

[1] Example text is written in `type writer font`.

remarkable properties will serve as an a posteriori justification.

The number of substrings of a text is $t_0(t_0+1)/2$ where $t_0$ denotes the text length. This expression quickly gets very large. The practical computation of $V$ is carried out using the *suffix tree* of $T$ which can be built in linear time and space complexity [12]. $k$ is then simply the number of nodes in this tree-like index structure [6].

The focus of this paper is not the value of $V$ for the whole text, but how $V$ develops if we read the text character by character and view $V$ as a function of $t$, the length of the text part read so far.

## 2.2 Exemplifying $V(t)$

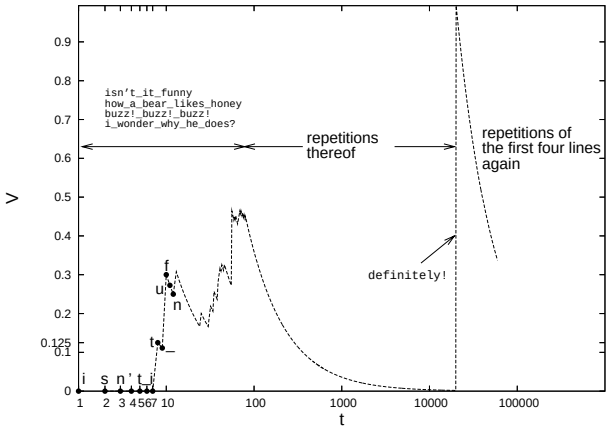


**Fig. 1:** *The development of $V$ for an example text set up to clarify the interpretation of $V$ as a measure of the level of repetitions. The scale on the x-axis is logarithmic.*

Fig. 1 shows the development of $V$ for an artificial example text of 60,149 characters: the first 20,000 characters are repetitions of the following four lines:

```
isn't_it_funny
how_a_bear_likes_honey
buzz!_buzz!_buzz!
i_wonder_why_he_does?
```

After this, new text (`definitely!`) is introduced, before the bear song repeats again until the end of the text.

For the first six text characters (`isn't_`) there is no repetition ($V = 0$). The seventh one (`i`) is a repetition of the first one. But only after the eighth character is read (`t`), does it have two different continuations (`s` and `t`). $V$ jumps to 1/8. The `t` itself is a repetition but the next character (`_`) is no new continuation and $V$ drops to 1/9. After the ninth character (`f`) is read, we have three substrings with different continuations (`i`, `t_` and `_`) and $V = 3/10$. The following `u` and `n` don't terminate any repetition, and $V$ drops again. In this way, $V$ follows a slow upward trend until the end of the four cited lines.

Nothing new is introduced now for nearly 20,000 characters. Since no repetition does ever terminate in this phase, $V$ drops steadily.

When this very long repetition is ended by the sudden appearance of `definitely!`, the situation changes radically. All at once we have nearly $20,000$ substrings with different continuations and $V$ jumps to a value close to 1 accordingly. After this interruption, the text gets repetitious again and $V$ drops for a second time.

# 3 Languages and corpora

$V(t)$ was compared for natural language texts from 20 languages. They belong to the three language families Indo-European, Dravidian, and Uralic. Their writing systems instantiate three different classes of writing systems.

## 3.1 The investigated languages

Regarding the genetic relations of the tested languages we refer to [1].

Fourteen Indo-European languages were investigated: The Slavic language Russian, the West Germanic languages English and German, the Romance language French, and the ten Indo-Iranic languages Assamese, Bengali, Gujarati, Hindi, Marathi, Oriya, Punjabi, Sinhala, Urdu, and Kashmiri (subclassified as Dardic).

Tamil, Kannada and Malayalam from the southern branch of the Dravidian language family were included, as was Telugu from the Telugu-Kui branch.

From the Finno-Ugric branch of the Uralic languages, the Finno-Saamic Finnish and the Ugric Hungarian were investigated.

## 3.2 The writing systems

The same text can come out completely different when written in different writing systems. Since the definition of $V$ is based on repetitions on the character level, the writing system used can be expected to affect the value of this quantity. To investigate this effect, the experiments have been performed on texts written with different scripts.

We adopt the classification of writing systems proposed in [4]. The authors classify scripts "with respect to how symbols relate to the sounds of the language"[4, p. 4]. The resulting classification of the scripts overlaps only in part with the genetic relations cited above.

### 3.2.1 Abugidas

"In an *abugida*, each character denotes a consonant accompanied by a specific vowel, and the other vowels are denoted by a consistent modification of the consonant symbols [...]."[4, p. 4].

Most languages spoken in the Indian language area use historically related abugidas. This applies to Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Oriya, Sinhala, Tamil and Telugu.

### 3.2.2 Abjads

"In a consonantary, here called an *abjad* [...] the characters denote consonants (only) [...]"[4, p. 4]

Some of the languages spoken in India use scripts based upon the Arabic script, the world's most widespread abjad. From the set of tested languages, this applies to Urdu, Kashmiri, and Punjabi. Urdu is an abjad following [4]. Regarding Kashmiri, see Section 3.2.3.

Punjabi is written in two different scripts: on the one hand in Gurmukhi (an abugida); on the other hand in the Perso-Arabic abjad. The corpus used for this investigation is written in Perso-Arabic.

### 3.2.3 Alphabets

"In an *alphabet*, the characters denote consonants and vowels"[4, p. 4].

German, English, Finnish, French and Hungarian use different variants of the Latin alphabet.

Russian is written with the Cyrillic alphabet.

The script used for Kashmiri is based on the Perso-Arabic abjad, but called an alphabet in [4]. I follow this classification.

## 3.3 The corpora

The corpora of the tested Indian languages are all part of the EMILLE corpus [18]. For each of these languages, I used between 2 and $20MB$ (that is approximately between 200,000 and 2 million tokens) of the written part of this corpus. Most of the data stems from various Indian dailies.

The German texts are taken from the online edition of the *Süddeutsche Zeitung* – a high quality German newspaper.

For English, a part of the Brown Corpus [17] was used.

For French [20], Russian [15], Finnish [14] and Hungarian [16], novels were used.

# 4 Experimental Results

Intuitively, there seems to be no reason for a uniform behaviour of $V(t)$ in different texts, let alone in different languages. On the contrast, it seems natural to expect changing levels of repetitiveness both within one text and between texts. The repetitiveness could probably depend on various factors such as subject, genre, author, the morphological structure of the language or the writing system.

Fig. 2 and Fig. 3 show the evolution of $V(t)$ for growing text sizes $t$. Fig. 3 is an enlargement of the central part of Fig. 2.

We can draw a set of observations from these figures:

**O1** For all investigated corpora $V$ converges towards a constant[2].

**O2** This constant is reached after as few as about 10,000 characters, that is after approximately three pages of text.

[2] The obvious jumps in most of the curves are addressed below.

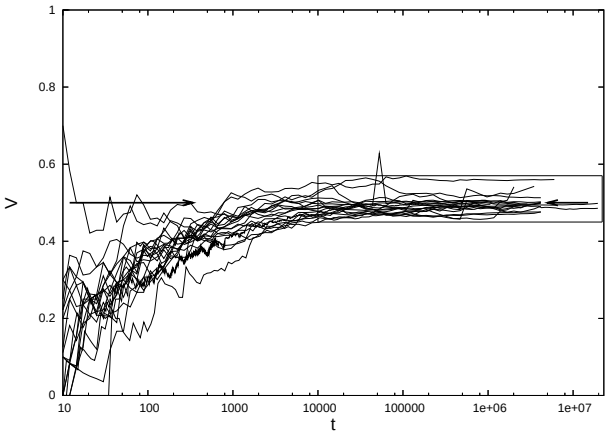


**Fig. 2:** *$V$ for all tested languages. Refer to the text for a list of languages. Fig. 3 enlarges the box in the middle and shows a label for each language. The characteristic value of $1/2$ is clarified by the bold arrows. The scale on the x-axis is logarithmic.*

**O3** For shorter text lengths, an average curve is easily discernible, although the convergence level is not yet reached.

**O4** The convergence level is compatible with 1/2.

We will henceforth summarise the uniform behaviour of the $V$-curves as described by the observations **O1** through **O4** under the term *$V$-convergence*. So far, $V$-convergence has been found in all tested natural language texts.

The $V$-curves of other texts show a much more diverse behaviour. A small set of examples of such texts is shown in Figure 4 (for comparison, $V(t)$ for the Rus-

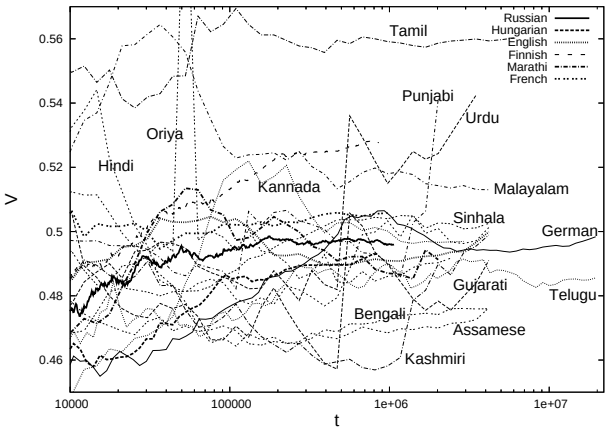


**Fig. 3:** *Enlarged middle part of Fig. 2.*

sian text is shown as curve R):

**1** A uniformly distributed random text, i.e. each character has the same probability of occurrence at each text position. Alphabet size is 3.

**2** `c` sources from the Linux 2.6.0 kernel. Generally, $V(t)$ for source code runs above 0.5 and shows a rather unpredictable behaviour.

**3** Random text generated as described in [3]. This elaborated language model was designed to emulate basic statistic characteristics of natural text such as mean word and sentence length.

**4** A random text which simulates the English character distribution.

**5** A uniformly distributed random text[3] with alphabet size 100. Compared with curve **1**, $V(t)$ looks rather different here. In general, for this class of texts, shape and height of $V(T)$ depend heavily on the alphabet size. This contrasts with the behaviour of natural language texts: although the set of symbols of abugidas (Section 3.2.1) is usually twice as large as for alphabets (Section 3.2.3), there seems to be no immediate impact on $V(t)$ (see Figure 3).

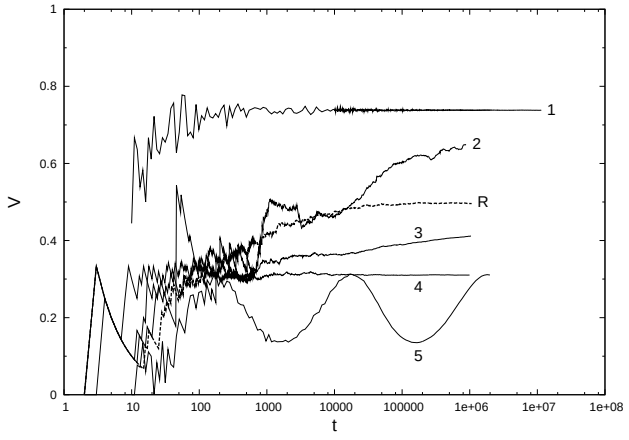


**Fig. 4:** *The $V$-curves of different kinds of text. See the text for a detailed description. The scale on the x-axis is logarithmic.*

On the grounds of the experiments reported here, the still highly speculative hypothesis can be formulated that $V$-convergence might be a universal and exclusive feature of natural language texts.

This hypothesis has to be thoroughly checked by testing many more texts from different languages, scripts, styles and epochs. So far, additional experiments with the Chinese LCMC corpus [19] were performed. This corpus exists in two different scripts, the traditional characters and the romanised transcript pinyin. The $V$ of the character version converges towards $0.27 \pm 0.02$, while the pinyin version shows $V$-convergence with $V$ approaching $0.52 \pm 0.01$. See also the discussion at the end of Section 5.

[3] The peculiar oscillations reflect the fact that first the bigrams repeat, then the trigrams, and so on.

**A Remark:** the bumps that can be seen for many of the corpora in Fig. 3 are due to longer repetitions in these corpora as exemplified in Fig. 1. For web based corpora, such as the EMILLE corpus [18], longer repetitions are hard to avoid, since cut and paste can easily multiply chunks of text or whole texts, especially if an online edition of a newspaper is used as the source.

The fact that longer repetitions in the text are reflected as bumps in its $V$-curve could be converted into a method for detecting artificial repetitions in large corpora, provided one is able to cope with the heavy memory consumption of suffix trees. This index structure is used for the technical implementation of the computation of $V(t)$ as mentioned in Section 2.1.

The novels and the German corpus don't show any bumps. For the novels, this smoothness can be expected, because long repetitions are naturally avoided. In the German corpus they occurred, but were carefully filtered out by a variety of ad-hoc heuristics.

# 5 The impact of randomisation

It is a special feature of the quantity $V$ that it is based on character strings, not on words. Accordingly, it measures repetitions both below and above word level. It is a natural question, which of these two kinds of repetitions contribute more to the value of $V$. To address this question, I separately randomised the internal structure of words and the sequence of words.

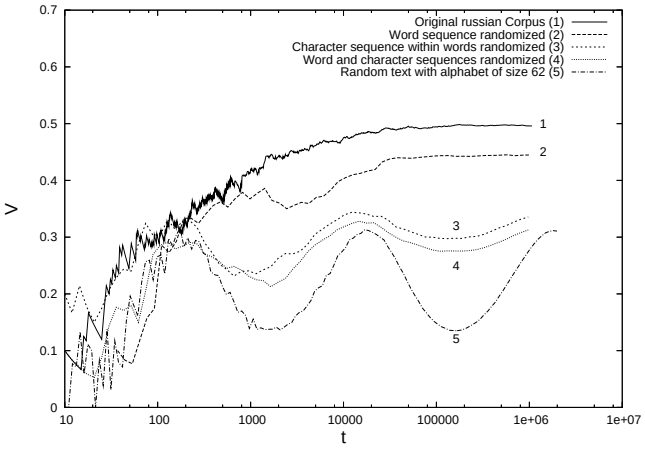


**Fig. 5:** *The impact of randomisation on $V(t)$. See the text for a detailed description.*

The starting point of the investigation was the original Russian corpus. The result of the different randomisations of the corpus is shown in Fig. 5:

**1** $V(t)$ for the original Russian corpus.

**2** The inner structure of words is left untouched, but their sequence is scrambled. $V(t)$ is considerably lowered.

**3** The characters of the words are randomised, while the word order is left untouched. Each word is replaced by a random character string. Equal surface forms are replaced by equal random strings

drawn from a 59 character alphabet. Each character had the same probability.

**4** combining the randomisation schemes for curves **2** and **3**.

**5** for comparison only: $V(t)$ for random text, drawn from a uniformly distributed alphabet of size 62.

Clearly, randomisation always lowers $V$. This is to be expected since a random string of characters and words can only result in random repetitions. Since the system of human language prescribes the reoccurrence of certain structures, we expect that a deliberate destruction of these structures will diminish the level of repetitions.

Randomising the internal structure of the words affects $V$ much more than only randomising word order. This shows that repetitions below word level contribute more to the value of $V$ than repetitions on a larger scale. This corresponds to another observation: $V$-convergence occurs in all tested texts written with scripts in which graphemes and phonemes correspond. This includes the (invented) pinyin script, but does not apply to the traditional Chinese script for which no $V$-convergence was found. Together, these two observations could be interpreted as a first hint that this phenomenon is rooted in the phonemic level of language.

# 6 Other text statistical regularities and constants

This section discusses how $V(t)$ and the phenomenon of $V$-convergence can be compared with other text statistical constants and regularities.

The best known such regularity is *Zipf's Law* [13]. It states that the most frequent word in any natural language text is twice as frequent as the second most frequent one and three times as frequent as the third most frequent one, and so on. Zipf's law roughly holds, except for the most frequent and the very infrequent words. But, as sketched in [9], and shown in more detail in [8], Zipf's Law is also valid for random text. This over-generality greatly reduces its significance: there's little value in knowing about a property which natural language text shares with noise. As pointed out in the discussion of Figure 4, $V$-convergence, on the other hand, could so far not be observed for random text, even if it simulates a natural character distribution or was designed to simulate the statistical features of natural language [3]. If it can be confirmed that $V$-convergence is a universal and exclusive feature of natural language text, we gain a strong tool to decide about the adequacy of statistical language models: if such a model is not able to reproduce $V$-convergence in its output, it cannot be said to mimic the structure of human language. This hurdle can be expected to be much higher for models which aim at modelling both words and their sequence. Models which reuse existing natural language words will have a lesser problem, as we know that the word sequence has a smaller impact on $V$ than the inner structure of the words themselves (see Section 5).

Besides *Zipf's Law*, a lot of lexicostatistic quantities were proposed to measure – for example – lexical richness or the productivity of word formation processes [2, 10]. Many of these text statistic quantities were proposed as constants, independent of text size. But it was shown that, in practice, these alleged constants tend to vary with text length [11]. Similarly, there is a class of models which try to capture these text length dependencies. Evert and Baroni [5], however, show that the predictive power of most of these models is low: the behaviour computed for small text sizes cannot be extrapolated to larger texts. In contrast to this, $V(t)$ converges very rapidly towards a fixed value around which it fluctuates only a little.

As can be seen from Fig. 1 and 4, different kinds of text can produce qualitatively diverse $V$-curves. In contrast, most lexicon based text statistical measures have only a few degrees of freedom. Consider Zipf's law as an example: it is usually depicted in a *Zipf plot*: starting with an ordered frequency list, the place in this list is shown on the x-axis, while the frequency is shown on the y-axis. This will always yield a monotonously decreasing function. The potential variability in $V(t)$ makes its uniformity in natural language text more surprising than the validity of Zipf's Law.

$V(t)$ is computed from the full character sequence of the text and is thus sensitive to structural changes on all levels. In contrast, the lexicostatistic quantities discussed in this section are usually derived from summary statistics such as the number of Hapax Legomena or the vocabulary size. Thus, they lose, from the start, most of the information contained in the full text: they remain the same if the text is replaced by a random sequence of random tokens, as long as these tokens have the same frequency distribution as the tokens of the original text.

As a consequence, none of the randomisation methods applied in Section 5 would have any effect on these word frequency based measures, since the statistics of the lexicon is left untouched.

This striking difference between lexicostatistic measures and constants, on the one hand, and $V$-convergence, on the other hand, effectively counters the argument that the latter might turn out to be an alternative manifestation of one of the former, for example of Zipf's Law.

All these features – its exclusive occurrence in natural language text, its higher sensitivity to structural changes of the text, its stable convergence and its richer structure – make $V(t)$ and its convergence towards $1/2$ much more informative and significant than any of the token frequency related models and constants.

# 7 Outlook

If $V$-convergence can be firmly established as a feature of natural language text, this would immediately raise two questions: why is the level of repetitions so very constant? It is clear that too many repetitions in language are bad: it's both boring and time consuming. On the other hand, if nothing ever repeats we have no chance of recognising known elements or

of regaining lost information: no understanding without repetition and no stable communication without redundancy. But why should repetitions be so evenly distributed? $V = 1/2$ seems to be some kind of optimum, but what does it optimise? The other question that would be raised is: what keeps $V$ this constant? What is the mechanism within the human language system that regulates repetitiveness?

But before all of these questions can gain real relevance, a second round of experiments is necessary: $V(t)$ has to be investigated for more texts – natural and non-natural – being as diverse as possible.

In order to get a clearer picture of $V$ and the phenomena surrounding it, the exact shape of this quantity will have to be measured carefully. One obvious questions is whether there is a significant deviation of the convergence point of $V(t)$ from 1/2 or not.

Another thrilling task ahead is to examine $V(t)$ for spoken corpora, maybe in phonetic transcription. Is $V$-convergence a phenomenon of written language or does it also occur in spoken language?

A related project [7] investigates the impact of stylistic differences, like authorship, on similar data.

## Acknowledgements

I thank my doctoral advisor Prof. Dr. Anke Lüdeling for her indispensable support, Prof. Dr. Klaus Schulz for giving me the opportunity of carrying out fundamental research, and Karsten Tabelow, Verena Harpe, Adriana Hanulíková and Anna McNay for critical comments and proof reading. Last but not least, I thank the anonymous referees for their valuable comments.

## References

[1] R. E. Asher and J. Simpson, editors. *The Encyclpedia of Language and Linguistics*. Pergamon Press, Oxford, New York, Seoul, Tokyo, 1994.

[2] R. H. Baayen. *Word frequency distributions*. Kluwer, Dordrecht, 2001.

[3] C. Biemann. A random text model for the generation of statistical language invariants. In *Proceedings of HLT-NAACL-07*, Rochester, NY, USA, 2007.

[4] P. Daniels and W. Bright. *The World's Writing Systems*. Oxford University Press, 1996.

[5] S. Evert and M. Baroni. Testing the extrapolation quality of word frequency models. In *Proceedings of Corpus Linguistics 2005*, 2006.

[6] F. Golcher. Statistische Aspekte von Suffixbäumen natürlichsprachiger Texte, Feb. 2005. Thesis for postgraduate studies at the University Munich (in German[4]).

[7] F. Golcher. A new text statistical measure and its application to stylometry. In *Proceedings of Corpus Linguistiscs 2007*, to appear.

[8] R. F. i Cancho and R. V. Solé. Zipfs law and random texts. *Advances in Complex Systems*, 5:1–6, 2002.

[9] W. Li. Random texts exhibit zipf's-law-like word frequency distribution. *ieeeit*, 38(6):1842–1845, 1992.

[10] A. Lüdeling and S. Evert. Linguistic experience and productivity: Corpus evidence for fine-grained distinctions. In *Proceedings of the 2003 Corpus Linguistics Conference*, Lancaster, 2003.

[11] F. J. Tweedie and R. H. Baayen. How variable may a constant be? measures of lexical richness in perspective. *Computers and the Humanities*, 32:323–352, 1998.

[12] E. Ukkonen. On-line construction of suffix-trees. *Algorithmica*, 14(3):249–260, 1995.

[13] G. K. Zipf. *Human Behavior and The Principle of Least Effort*. Hafner Publishing Company, New York, London, 1949.

## Corpora

[14] F. Bremer. Koti[5] [online]. July 2005 [cited 03/04/07]. Available from World Wide Web: `http://www.gutenberg.org/dirs/etext04/8phnm10.txt`. Finnish by Alma Suppanen, Original published in 1839.

[15] F. Dostoevsky. Crime and Punishment (in Russian) [online]. 2004. Available from World Wide Web: `http://lib.ru`. Biblioteka Maksima Moshkova.

[16] G. Gèza. Egri Cssilagok. PDF of unknown origin, 1899.

[17] H. Kuĉera and W. N. Francis. *Computational Analysis of Present-Day American English*. Brown University Press, Providence, RI, USA, 1967.

[18] A. McEnery, P. Baker, R. Gaizauskas, and H. Cunningham. EMILLE: building a corpus of South Asian languages. *Vivek, A Quarterly in Artificial Intelligence*, 13(3):23–32, 2000.

[19] A. McEnery, Z. Xiao, and L. Mo. Aspect marking in english and chinese: Using the lancaster corpus of mandarin chinese for contrastive language study. *Literary and Linguistic Computing*, 18(4):361–378, 2003.

[20] M. Proust. Du Côté de Chez Swann [online]. 2001 [cited June 20 2006]. Available from World Wide Web: `http://www.gutenberg.org/etext/2650`. Projekt Gutenberg – plain text version.

[4] Translated title: *Statistical aspects of suffix trees of natural language texts*

[5] *Home*, Swedish original: *Hemmet*