Rank-frequency Relation & Type-token Relation: Two Sides of the Same Coin

Jiří Milička

Institute of Comparative Linguistics, Charles University, Prague
Jana Palacha 2, Praha 1, 116 38

Milicka@centrum.cz

Abstract. This paper shows that type-token relation, hapax-token relation and, generally, relation between types of certain frequency and tokens can be computed from the rank-frequency relation or from any type of frequency distribution and that type-token relation can be computed from the hapax-token relation. This paper shows that there is no need for any approximation or assumptions and that the formulae can be derived purely algebraically. The second part of the paper observes that, for a very large corpora, the ratio between the number of hapax legomena and types converges to a constant *Z*; *Z>0*. Under this assumption an approximation is built that enables us to predict type-token relation and other aforementioned relations from the single parameter *Z*. This approximation is only valid for very large corpora. As the last chapter shows, this assumption implies that for an infinitely increasing number of tokens, the number of types increases beyond any limit.

Keywords: Type-token relation, hapax-token relation, rank-frequency relation, words frequency distribution.

1 Introduction

Type-token relation (TTR) and rank-frequency relation (RFR) are ones of the most popular ways of quantification of a text. They are used in empirical linguistics, NLP, literary theory etc. Many approximative models of these relations have been introduced since the dawn of quantitative linguistics. The most famous ones are Herdans Law for TTR (mostly referred to as Heaps' Law) and Zipf or Zipf-Mandelbrot Law for RFR.

In the first two chapters of this paper we won't see those relations through the prism of any approximation, on the contrary, the first chapter shows, that mere means of algebra are sufficient to transform any measured distribution of types (and thus RFR) into TTR curve, or into hapax legomena – token relation (T_1TR), or dis legomena – token relation (T_2TR), or any other relation between the number of types of certain frequency and the number of tokens (T_gTR). In the second chapter we use these formulae to derive a formula that exactly transforms T_1TR curve into TTR curve, and introduce some other formulae related to T_1TR and T_gTR.

The next chapter is based on these formulae and discusses consequences of an empirical observation that the ratio between the number of hapax legomena and the number of types asymptotically tends to a constant larger than zero. This works for all natural texts, even very large corpora.

2 Computation of TTR from a frequency distribution of types (or RFR)

The following formulae, which enable us to compute TTR, T_1TR and generally T_gTR of a text from a mere frequency distribution of types in a text, were derived 5 years ago [4]. In those days, the idea that it must be possible to compute TTR from RFR was quite common among researchers (e.g. [3]) but none of them approached this task without any assumptions. Dieter Müller even states, that TTR cannot be derived from the general distribution without an approximation. [1]

But it can. The formula transforming RFR (or any absolute frequency distribution of types, Zipfian or otherwise) into TTR is the following one[2]:

$$V(N) = \sum_{i=1}^{M} \left(1 - \frac{(d-N)!\,(d-f_i)!}{d!\,(d-N-f_i)!}\right) \tag{1}$$

We are able to compute T_1TR from RFR using the following one[3]:

$$V_1(N) = \sum_{i=1}^{M} \frac{\frac{N(d-N)!}{(f_i-1)!\,(d-N-f_i+1)!}}{\frac{d!}{f_i!\,(d-f_i)!}} \tag{2}$$

And in the most general case, we can compute the relation between types of a certain frequency and tokens (T_gTR) from a types distribution (RFR) according to this formula[4]:

$$V_g(N) = \sum_{i=1}^{M} \frac{\frac{N!}{g!\,(N-g)!}\frac{(d-N)!}{(f_i-g)!\,(d-N-f_i+g)!}}{\frac{d!}{f_i!\,(d-f_i)!}} \tag{3}$$

[1] "The general case of a vocabulary *V* with arbitrary type probabilities w_j requires an approximation" [5, page 204].

[2] $V(N)$ represents the number of types after measuring N tokens, i is a control variable that represents order of a type in the lexicon, f_i is the number of occurrences (absolute frequency) of the type in the text, M is the total number of types in the text, d is the total number of tokens in the text.

[3] $V_1(N)$ represents the number of hapax legomena.

[4] $V_g(N)$ is the number of types represented by g tokens.

The main idea on which the formulae are based is that (technically), although we cannot make all permutations of a text and measure TTR etc. for these permutations and average them, we can simulate this process by means of algebra.

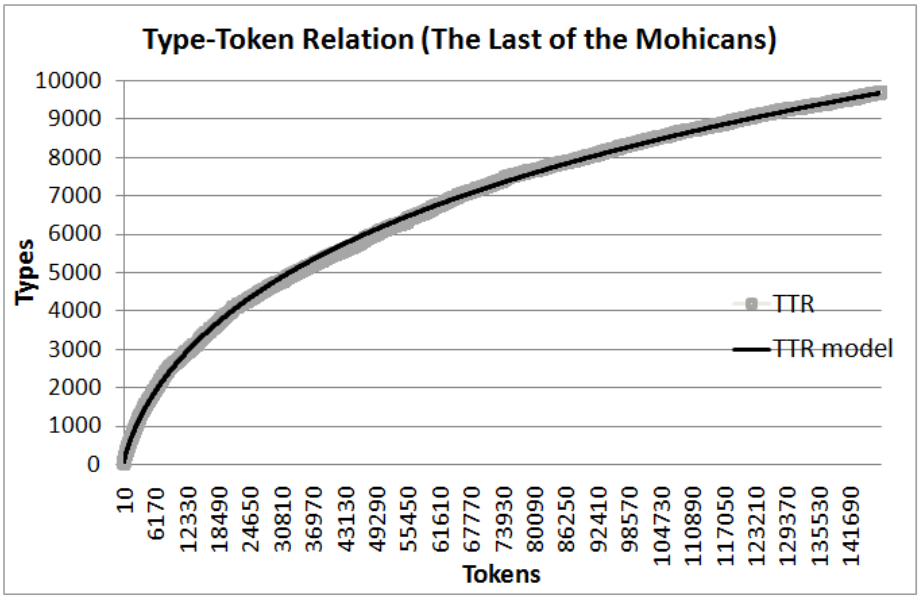

Fig. 1. Illustration of the usage of the model. Type-token relation measured and computed for The last of the Mohicans by J. F. Cooper [9].

Fig. 2.

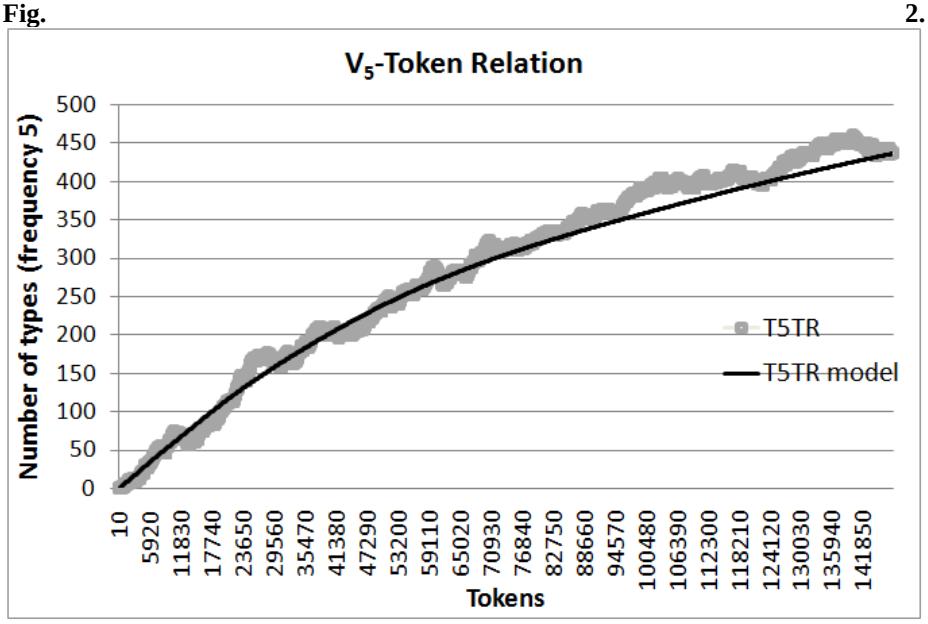

Fig. 3. Illustration of the usage of the model. The relation between types with frequency 5 and tokens, measured and computed for The last of the Mohicans by J. F. Cooper [9].

The fact that the formula corresponds to the quantities measured on natural texts tells us that the text is a homogenous one, i.e. the writer chose words similarly as if he chose it randomly from a multiset of words.

How we arrived at these formulae can be found in [4]. They were introduced here because the next chapter is based on them.

3 Computation of TTR from T_1TR

The following formula is based on the previous ones and thus is also distribution-independent. It transforms T_1TR into TTR and vice versa exactly (without any approximation). The underlying idea is very simple:

Consider a pack consisting of the cards with these suits:

♥ ♠ ♣ ♥ ♦ ♣ ♥ ♥

And imagine that you take away one card randomly. The probability that the number of suits in the pack decreases is equal to the number of suits that are only once in the pack divided by the number of cards[5].

[5] Given a multiset A and a multiset B; $B = supp.A$ and a multiset C; $C = B \ominus supp.(A \ominus B$. The probability that the cardinality of a multiset B decreases when decreasing cardinality of A by one is equal to cardinality of C divided by cardinality of A.

$$V(N) - V(N-1) = \frac{V_1(N)}{N} \qquad (4)$$

Similar formula has been published by Baayen (p. 115, [1])

We can also derive the formula directly from the formulae described in the previous chapter and thus complete them into one consistent framework.

$$\sum_{i=1}^{M}\left(1 - \frac{(d-N)!\,(d-f_i)!}{d!\,(d-N-f_i)!}\right) - \sum_{i=1}^{M}\left(1 - \frac{(d-N+1)!\,(d-f_i)!}{d!\,(d-N+1-f_i)!}\right) =$$

$$= \frac{1}{N}\sum_{i=1}^{M}\frac{\mathrm{N}\,(d-N)!}{\frac{(f_i-1)!\,(d-N-f_i+1)!}{d!}}$$

$$\sum_{i=1}^{M}\frac{(d-N)!\,f_i(d-f_i)!}{(d-N-fi+1)!} = \sum_{i=1}^{M}\frac{(d-N)!\,f_i(d-f_i)!}{(d-N-fi+1)!} \qquad (5)$$

The whole proof you can find on www.milicka.cz/kestazeni/beograd/dukaz1.pdf.

Now, a measured T_1TR curve can be transformed into TTR easily:

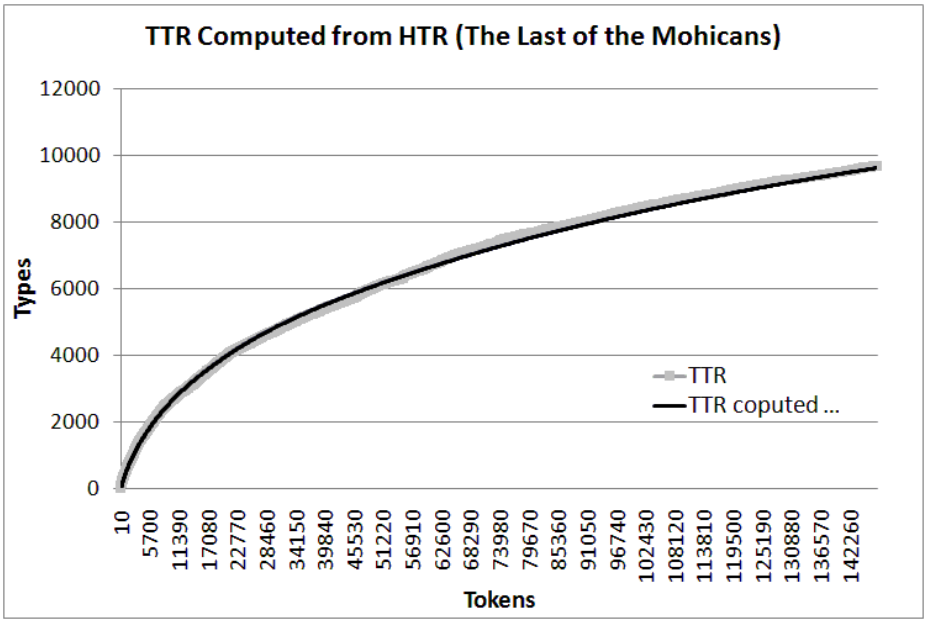

Fig. 3. Type-token relation measured on a text (The Last of the Mohicans by J. F. Cooper [9]) and the same curve computed from hapax-token relation measured on the same text.

Another form of the formula is[6]:

$$\frac{V(N)}{V(N-1)} = \frac{N}{N - Z(N)} \qquad (6)$$

An inverse transformation (TTR into T_1TR) is not easy when using real data, because real world difference between V(N) and V(N-1) is heavily influenced by random deviations. However, this does not mean that the formula is not valid and that we cannot use it to derive other formulae.

The following formula is based on the same idea as the previous one and we can also prove it using the combinatorial model[7]. It expresses the exact relation between the number of types represented by g tokens and the number of types represented by $g + 1$ tokens.

$$V_g(N) - V_g(N-1) = \frac{gV_g(N) - (g+1)V_{g+1}(N)}{N} \qquad (7)$$

The same formula in another form:

$$V_g(N) = \frac{(g-1)V_{g-1}(N) - N\left(V_{g-1}(N) - V_{g-1}(N-1)\right)}{g} \qquad (8)$$

And the most general one:

$$V_{g(N)} = N\frac{V(N) - V(N-1) - \sum_{i=1}^{g-1}\left(V_{i(N)} - V_{i(N-1)}\right)}{g} \qquad (9)$$

Because of the random deviations, these formulae are not very useful directly for the real life data, but we will need them in the next chapter.

4 The approximation

In this chapter we expand outside pure algebra and take into account an assumption that (for large monolingual corpora) the ratio between the number of hapax legomena and the number of all types converges to a constant larger than zero.

[6] Where $Z(N) = V_1(N)/V(N)$.

[7] Proof available on www.milicka.cz/kestazeni/beograd/dukaz2.pdf

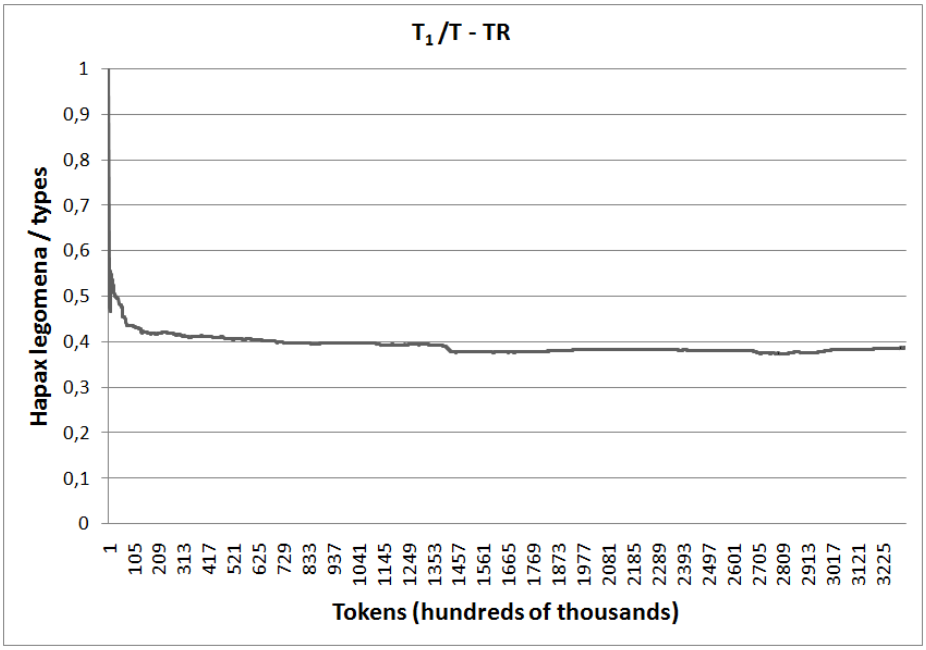

Fig. 4. The ratio between the number of hapax legomena and the number of all types converges (for Arabic corpora CLARA [10] and CLAUDIA [11]) to cca 0.38.

This assumption was widely discussed in [2], where Cvrček even claims that after an initial drop the ratio slightly increases (measured for wordforms and lemmas in large European languages corpora). This assumption allows us to modify formula 6, where Z(N) is a function, resulting in the following formula:

$$\frac{V(N)}{V(N-1)} = \frac{N}{N-Z} \quad (10)$$

where Z is a constant[8], at least for very large corpora consisting of hundreds of millions of tokens. Here is the non-recurrent form of the formula (using Pochhammer's symbol).

$$V(N) = \frac{V(M)(1)_{(N-M)}}{(1-\mathrm{Z})_{(\mathrm{N-M})}} \quad (11)$$

Parameter M expresses the number of tokens by which we assume that the hapax-type ratio reached its constant value. V(M) is the number of types after measuring M tokens. It is our initial position, starting point.

We can use the formula in real life to predict TTR for a very large amounts of data, simply by measuring TTR and T_1TR until Z(N) is satisfactorily stable. Then we

[8] $Z = lim_{N\to\infty} V_1(N)/V(N)$

consider the reached token as an initial position *M*, so far measured number of types as initial V(M) and from this initial position further we can model the growth of the number of types by putting the figures into the following formula:

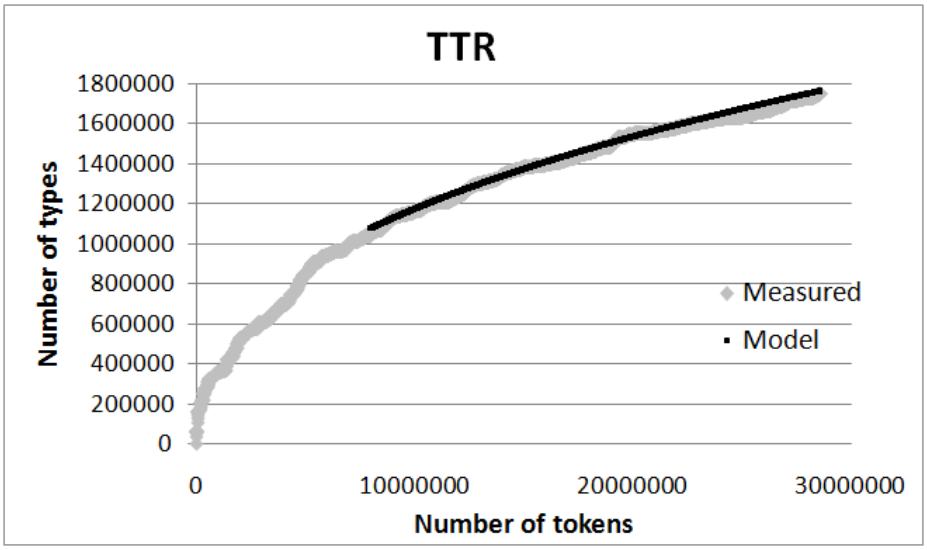

Fig. 5. Illustration of usage of the model. Type-token relation measured and computed for the CLAUDIA corpus [10].

It can be interesting for some linguists that this formula tells us that if the constant *Z* is larger than zero then the number of types grows beyond any limit.

$$\lim_{N\to\infty} \frac{V(M)(1)_{(N-M)}}{(1-\mathrm{Z})_{(\mathrm{N-M})}} = \infty \quad (12)$$

The similar formula is also valid for T_1TR[9]:

$$\frac{V_1(N)}{V_1(N-1)} = \frac{N}{N-Z} \quad (13)$$

And by mathematical induction[10] we can prove that the following similar formula is also valid for the growth of the number of any types of any other frequency.

$$\frac{V_g(N)}{V_g(N-1)} = \frac{N}{N-Z} \quad (14)$$

[9] For the proof see www.milicka.cz/kestazeni/beograd/dukaz3.pdf
[10] For the proof see www.milicka.cz/kestazeni/beograd/dukaz4.pdf

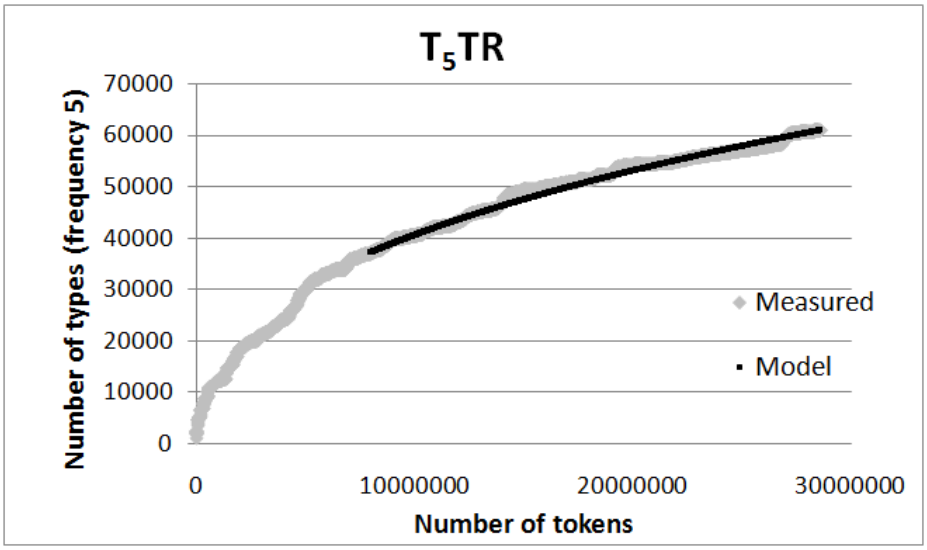

Fig. 6. Illustration of the usage of the model. The relation between types with frequency 5 and tokens, measured and computed for the CLAUDIA corpus [10].

From these formulae we can arrive[11] at the final formula that enables us to calculate the number of types of a certain frequency (V_g) using the constant Z as the only parameter.

$$\frac{V_g}{V_{g-1}} = \frac{g - Z - 1}{g} \quad (15)$$

This formula enables us to transform the number of types into frequency density function, which is possible to transform into distribution of frequencies of types or RFR.

[11] For the proof see www.milicka.cz/kestazeni/beograd/dukaz5.pdf

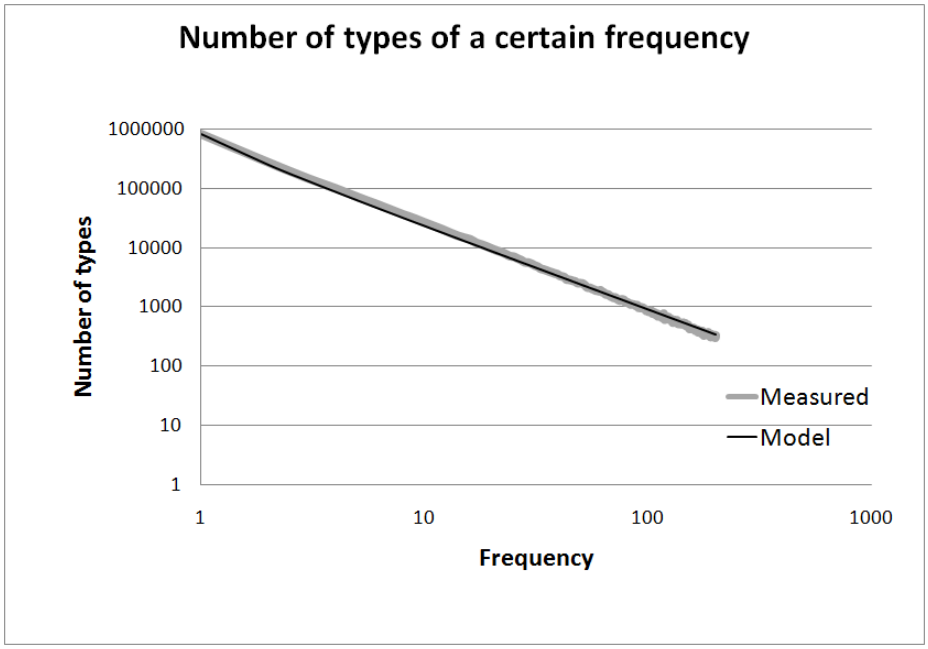

Fig. 7. Dependency of the number of types with certain frequency on the frequency (log-log). Measured on CLARA [10] and CLAUDIA[11].

5 Conclusions

1. RFR, TTR, T_1TR, T_gTR and also the average frequency of types etc. are one phenomenon. If two texts substantially differ in one of these quantities, they would be also different in other ones.
2. We can exactly calculate TTR, T_1TR and T_gTR form RFR or a distribution of types. We can also exactly calculate TTR from T_1TR. For inverse relations we still need an approximation.
3. A good approximation of V_1/V would enable us to calculate TTR, T_gTR, RFR etc.
4. Even if we consider V_1/V to be a constant, we can successfully model TTR, T_gTR, RFR etc. for very large corpora (because V_1/V seems to converge to a constant). This constant is the only parameter for all of these approximations.
5. V_1/V converging to a constant larger than zero implies that the number of types (and the number of types of a certain frequency) will increase to infinity as the number of tokens increases to infinity.

6 References

1. Baayen, R. H.: Quantitative aspects of morphological productivity. In: G. E. Booij and J. van Marle (eds), **Yearbook of Morphology 1991**, pp. 109–149. Dordrecht (1992).
2. Cvrček, V.: How large is the core of language? (In press).
3. Leijenhorst,D. C. van – Weide, Th. P. van der: A formal derivation of Heaps' Law. In: Information Sciences 170, pp. 263–272. New York (2005).
4. Milička, J.: Type-token & Hapax-token Relation: A Combinatorial Model. In: **Glottotheory 2/1**, pp. 99–110. Trnava (2009).
5. Müller, D.: Computing the Type Token Relation From the A Priori Distribution of Types. In: **Journal of Quantitative Linguistics, Vol. 9, No 3**, pp. 193–214 (2002).
6. Syropoulos, A.: Mathematics of Multisets. In: **Multiset Processing**, pp. 347–358.: Springer-Verlag, London (2001).
7. Wimmer G.: The type-token relation. In: R. Köhler (ed.): **Quantitative Linguistics. An International Handbook**, pp. 361–368, (2005).
8. Wyllys, R. E.: Empirical and Theoretical Bases of Zipf's Law. In: **Library Trends 30(1) no. 53–64**, pp. 53–64, (1981).

7 Corpora

9. CLARA (cca 40M tokens) Synchronic Arabic Corpora (MSA)
10. CLAUDIA (cca 300M tokens) Diachronic Arabic Corpora
11. Cooper, J. F.: The Last of the Mohicans