CDMTCS
Research
Report
Series

Stratification in Texts

Gabriel Altmann[1], Ioan-Iovitz Popescu[2], Dan Zotta[2]
[1]Lüdenscheid and [2]Bucharest

CDMTCS-433
February 2013

Centre for Discrete Mathematics and
Theoretical Computer Science

Glottometrics 25, 2013,85 - 93

Stratification in texts

Gabriel Altmann, Lüdenscheid
Ioan-Iovitz Popescu, Bucharest[1]
Dan Zotta, Bucharest

Abstract. Stratification in texts is a process analogous to those in nature and culture. Though one cannot identify the individual strata in every case, it is possible to show the rise of this phenomenon in mathematical terms and apply the resulting formulas to examples from textology and music. It allows also to study the evolution of a writer, text sort, language or music.

Keywords: *stratification, text, music, differential equations*

Stratification is a property inherent to all material things. Modern science, especially physics, has shown it in innumerable cases and the process of discovery continues incessantly. But even human artefacts have strata. Some of them are created by concept formation in order to give us orientation and a basis for analysis, other ones are necessary for the artificial thing itself in order to be considered as such, e.g. colour or grey strata for pictures and paintings; pitch height, length, intensity, rhythm and colour for music; words, blanks, punctuation for writing; segmental and suprasegmental strata for spoken language, etc. Long time ago linguists stated that an utterance is stratified, even if is it written: The text is not a homogeneous mass and even its simple understanding requires a multistratal analysis which is automatized in the mother tongue and must be learned laboriously in foreign ones. Strata like sentence, clause, phrase, word, morpheme, syllable, phoneme are taught even in the school and they have the agreeable property that each stratum is linked with the neighbouring (higher or lower) stratum by means of Menzerath's law. Though this is a stochastic law, its existence contributes to the good conscience of linguistics to be a science just like its great sister, the physics.

But it would be foolish to suppose that our way ends at this point. There are at least three directions in which we can continue our way of stratification research. The first is the zone between text and its components. There are some purposefully created layers like chapters, paragraphs, acts in the stage play, decided by the author; other ones have been discovered and can be captured only analytically: up to now there is the "hreb" or sentence aggregate discovered by Hřebíček (1997) represented by all sentences of a text containing a synonym, a reference or some other identifying semantic connection between sentences; and the motif discovered by Köhler (2006, 2008a,b) consisting of non-decreasing sequences of some measured entities. The motif is a formal entity, hreb is rather a semantic one.

The second possibility is the classification of different entities in many different subclasses - a speciality and final aim of qualitative linguistics: there are parts-of-speech, grammatical categories, different types of morphemes, phrases, clauses, sentences, i.e. even within one class - which are merely Menzerathian chain-links in the hierarchy -, there are different substrata that can be identified formally or semantically. Though the author may select them deliberately, it would be very courageous to suppose that (s)he does not act in agreement with a law. One of such laws is e.g. Zipf's law in all its forms.

[1] Address correspondence to: Ioan-Iovitz Popescu, e-mail: iovitzu@gmail.com.

A third research possibility is the investigation of the number of sub-strata that occur within one stratum. An analogy with nuclear physics or microbiology is evident. We "open" the atom (being an element of a stratum) to see whether and what kinds of entities are in its interior; we open the DNA to see what it consists of. In linguistics, we arrived at a point at which we can at least state *how many* substrata are contained in a homogeneous stratum, e.g. that of words. We can, of course, state the frequency of word classes and see that synsemantics are more frequent than autosemantics, that short words are more frequent than long words but this all are properties constructed conceptually by us and follow some laws known from synergetic linguistics. But even these classes are combined in such a way that no grammar or semantics can approach them. The substrata may arise stepwise: by change of theme, by pauses in writing, by the development of the story, etc., but they can also be eliminated: the author may correct the text, the editor may strive for uniformity, etc. The reader/hearer need not even perceive a difference and most probably none of these text creators (writer, editor, reader) is conscious of something like strata in text.

The discovery and identification of strata in text - with whatever unit - is a problem for the far future. Though in stage plays there is a manifest stratification represented by different persons, other kinds are not easy to be identified. In some other domains of language it is easier to find strata, for example in the monolingual dictionary where each word is defined in terms of words which have a more general meaning. E.g. a "revolver" is a "weapon"; the weapon is an "instrument"; the instrument is an "artefact"; the artefact is a "thing". In this way one obtains strata of generality. Besides, it is evident that the more general the meaning, the fewer words are contained in the stratum. In the same way one can obtain strata of concreteness-abstractness, emotionality, metaphor, imagery, dogmatism, etc. known from psycholinguistics.

Nevertheless, there is a possibility of tracing down at least the existence of strata and their number in text using a mathematical reasoning. Unfortunately, it must be applied for each linguistic entity separately: if there is stratification in the vocabulary of the text, it need not exist e.g. for sentence length. In the second stage of the research it will also be necessary to substantiate the existence of strata linguistically.

We start from the following assumptions: The writer begins to write. At a certain (unknown) point in text he changes his strategy concerning certain units and continues with a slightly different strategy. Then somewhere he changes again to a new strategy that means, he performs a change of the change. In mathematical terms, the first change is $dy/dx = y'$; the change of this regime means simply a new change, i.e. $d^2y/dx^2 = y''$, etc. It is a matter of empirical fact that the function y and its derivatives obey a linear relationship, as will be shown in continuation.

Let us model a linguistic phenomenon which can be ranked, scaled or weighted. If the values converge to a constant (e.g. absolute frequencies converge to 1, relative frequencies to 0), we can always use the approach

$$(1) \qquad f(x) = C + y(x),$$

C being a real positive constant.

If we suppose the existence of stratification and restrict ourselves to two strata, we may express this assumption by

$$(2) \qquad y(x) = A_1 exp(k_1 x) + A_2 exp(k_2 x)$$

used successfully to rank-frequency sequences proposed as an alternative to Zipf’s law which does not capture stratification (cf. Popescu, Altmann, Köhler 2010). The derivatives of (2) are

$$\begin{aligned} y' &= A_1k_1exp(k_1x) + A_2k_2exp(k_2x) \\ y'' &= A_1\, k_1^2exp(k_1x) + A_2\, k_2^2exp(k_2x). \end{aligned} \tag{3}$$

From (2) and (3) we have the following differential equation

$$y'' - (k_1 + k_2)y' + (k_1\, k_2)y = 0 \tag{4}$$

where $k_1 \neq k_2$ are real numbers. Denoting further by

$$p = -\,(k_1 + k_2)$$

$$q = (k_1k_2)$$

we get the standard form of the 2nd order linear homogeneous ordinary differential equation with constant coefficients

$$y'' + py' + qy = 0. \tag{5}$$

Conversely, let's start from this equation

$$y'' + py' + qy = 0$$

where p and q are real numbers, and look for a solution

$$y = exp(kx).$$

Inserting it into the above equation we have

$$(k^2 \; + pk + q)exp(kx) = 0$$

or, because $exp(kx)$ is never zero, we obtain the so called *characteristic equation*

$$k^2 \; + pk + q = 0$$

with the *discriminant*

$$\Delta = p^2 - 4q$$

If $\Delta > 0$, the characteristic equation has two real and distinct solutions, k_1 and k_2, given by

$$k_1 = (-p + \surd\Delta) / 2$$
$$k_2 = (-p - \surd\Delta) / 2,$$

hence the corresponding solution of the considered differential equation is

$$y(x) = A_1 exp(k_1 x) + A_2 exp(k_2 x)$$

with A_1 and A_2 to be determined from initial conditions. Obviously,

$$p = -(k_1 + k_2)$$

$$q = k_1 k_2.$$

To conclude, the fitting function consisting of two exponential components represents the solution for the case Δ > 0, of the 2nd order linear homogeneous ordinary differential equation with constant coefficients, see more, for instance, at http://www.efunda.com/math/ode/linearode_consthomo.cfm

The generalization is straightforward: the fitting function consisting of *n* exponential components represents the solution of the *n*th order linear homogeneous ordinary differential equation with constant coefficients, for the case when all solutions of the characteristic equation are real and distinct numbers.

The above solution of the stratification problem has the advantage of telling us the number of strata of the given unit in the given text (cf. Popescu, Altmann, Köhler (2010); Popescu, Čech, Altmann (2011); Popescu, Mačutek, Altmann (2009); Popescu, Martináková-Rendeková, Altmann (2012)). However, it does not enable us to identify the strata.

Take as an example the word form frequency in Goethe's poem *Erlkönig* ranked in decreasing order as shown in Table 1.

Table 1
Ranked word form frequencies in Erlkönig by Goethe

x	f_x	x	f_x	x	f_x	x	f_x
1	11	32	2	63	1	94	1
2	9	33	2	64	1	95	1
3	9	34	2	65	1	96	1
4	7	35	2	66	1	97	1
5	6	36	2	67	1	98	1
6	6	37	2	68	1	99	1
7	5	38	2	69	1	100	1
8	5	39	2	70	1	101	1
9	4	40	1	71	1	102	1
10	4	41	1	72	1	103	1
11	4	42	1	73	1	104	1
12	4	43	1	74	1	105	1
13	4	44	1	75	1	106	1
14	4	45	1	76	1	107	1
15	4	46	1	77	1	108	1
16	3	47	1	78	1	109	1
17	3	48	1	79	1	110	1
18	3	49	1	80	1	111	1

19	3	50	1	81	1	112	1
20	3	51	1	82	1	113	1
21	3	52	1	83	1	114	1
22	2	53	1	84	1	115	1
23	2	54	1	85	1	116	1
24	2	55	1	86	1	117	1
25	2	56	1	87	1	118	1
26	2	57	1	88	1	119	1
27	2	58	1	89	1	120	1
28	2	59	1	90	1	121	1
29	2	60	1	91	1	122	1
30	2	61	1	92	1	123	1
31	2	62	1	93	1	124	1

If we fit the data with a function having a sum of three exponential functions in its expression, that is with

(6) $f(x) = 1 + A_1 exp(k_1 x) + A_2 exp(k_2 x) + A_3 exp(k_3 x)$,

we obtain the results presented in Figure 1 with the determination coefficient $R^2 = 0.9824$.

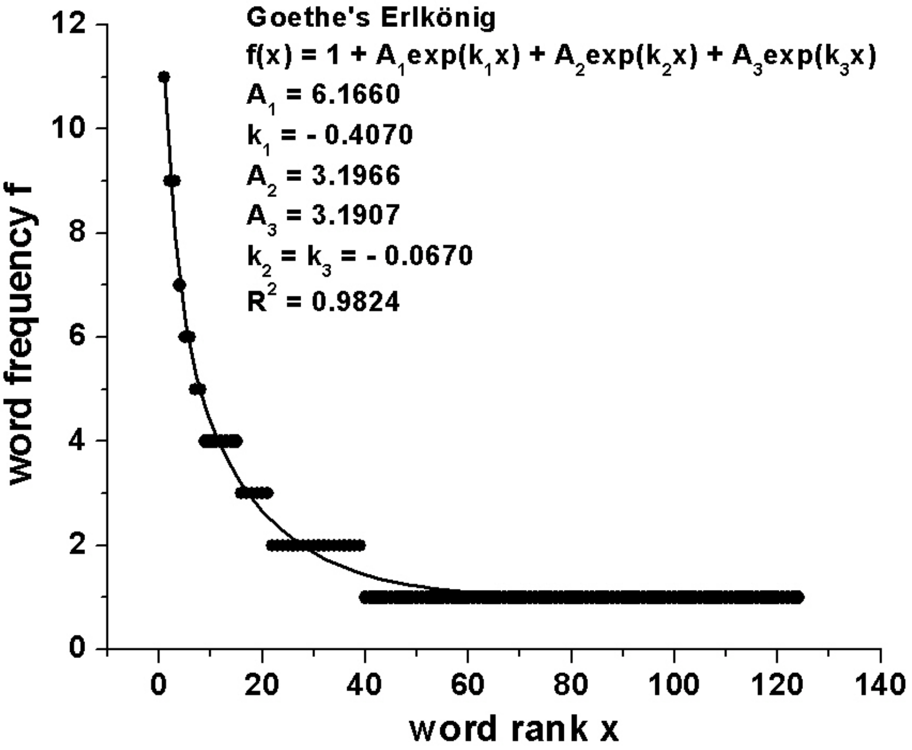

Figure 1. Fitting the word rank-frequencies in *Erlkönig* by Goethe with a function of type (6) indicates two strata

As can be seen, the parameters in the exponent k_2 and k_3 are equal hence we can omit one component and add the corresponding multiplicative constants $A_2 + A_3$. One obtains finally

$f(x) = 1 + 6.1160_1 exp(-0.4070x) + 6.3872 exp(-0.0670x)$

We can conclude that concerning word forms the poem has two strata. The function can be enlarged to more components - following from the differential equation of *n*-th order - but in case that some of the parameters yield non-realistic values, e.g. too great ones, one should omit them as outliers. It is to be noted that using the exponential function with one component we obtain still very good fitting results ($R^2 = 0.9648$) but we do not learn how many components there are. Hence the above method should be started always with several components. The next (qualitative) step would be the *identification* of the two strata, but this is more or less a philological affair.

This technique has been successfully used in many cases cf. e.g. Tuzzi, Popescu, Altmann, (2010: Ch. 5.1, 5.2), Nemcová, Popescu, Altmann (2010), Fan, Altmann (2010), Beliankou, Köhler (2010), Sanada, Altmann (2009), Laufer, Nemcová (2009), Kelih (2009), Knight (2013), etc. It is to be noted that this approach does not yield a "text model", it is merely a means to find the number of strata. There are always functions which would yield better fittings but their interpretation is quite different.

Let us consider some musical examples in which we found different stratifications.

Consider first the pitch rank-frequencies in Stravinsky's *The Firebird Suite*. Beginning with three components we obtain the result presented in Figure 2. As can be seen, all parameters in the exponent are identical, hence there is only one stratum and the computed rank-frequencies abide by $f_x = 1 + 265.9074exp(-0.0686x)$ where the parameters A_i were summed up.

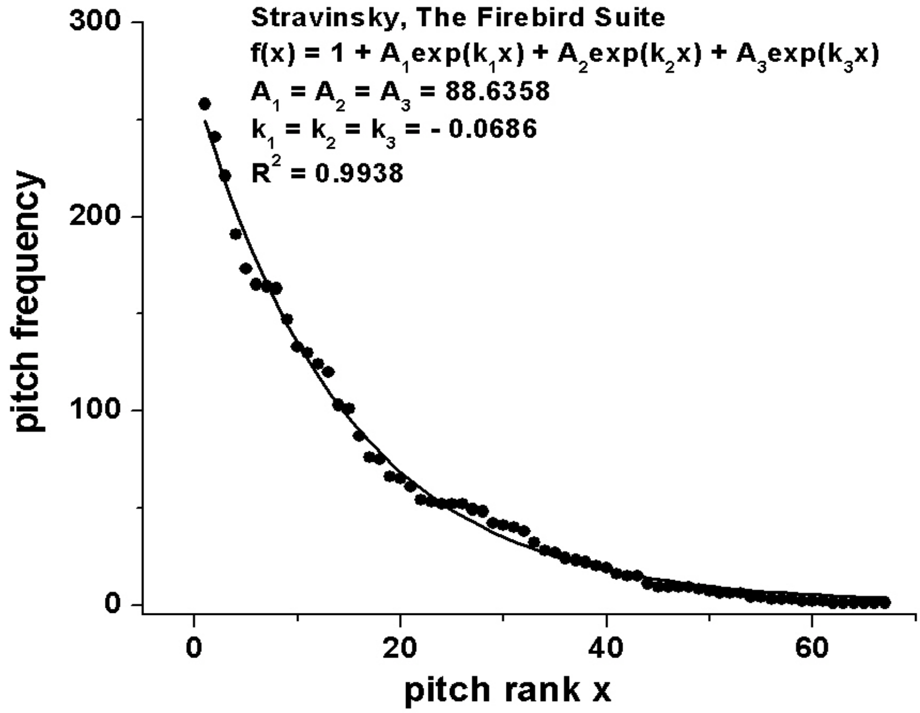

Figure 2. Fitting the pitch rank-frequencies in Stravinsky's *The Firebird Suite* with a function of type (6) indicates a single stratum.

In Beethoven's *Sonata No. 5*, presented in Figure 3, we find two strata because $k_2 = k_3$, hence $f_x = 1 + 93.1319exp(-0.7054x) + 446.3417exp(-0,0594x)$.

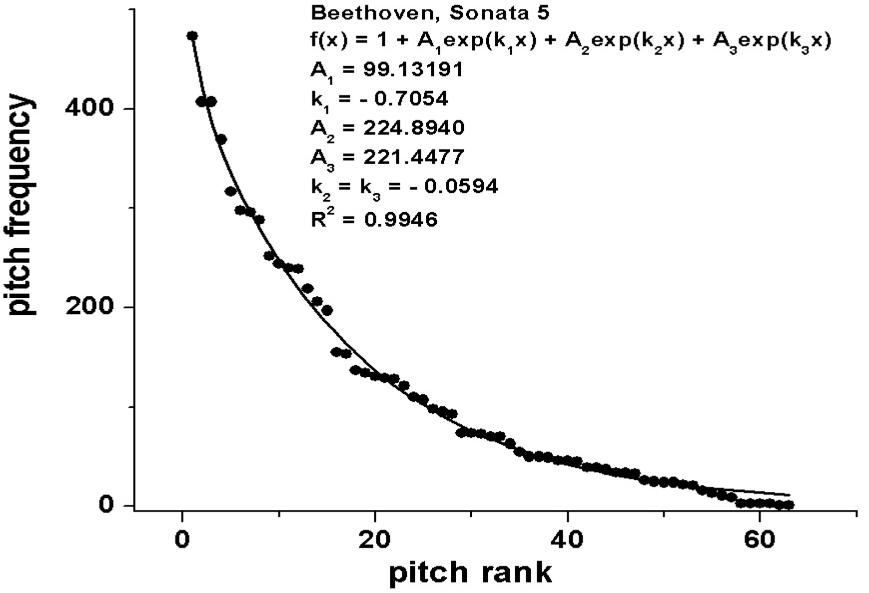

Figure 3. Fitting the pitch rank-frequencies in Beethoven's S*onata 5* with a function of type (6) indicates two strata.

A critical case is Mozart's *Sonata A major K.331* presented in Figure 4 indicating three strata but actually, there are only two strata because the excessively high multiplicative constant A_1 = 16869.3891 value corresponds to an outlier. If we compute directly two strata, we obtain f_x = 1 + 822.0111*exp*(-0.0853*x*) + 2322.1284*exp*(-2.2435*x*) with R^2 = 0.9942. But even here we have still A_2 = 2322.1284 which is more than twice the observed f_1 = 1002. If we consider it an outlier, we obtain the monostratal fitting in form f_x = 1 + 923.0682*exp*(-0.0951*x*) with R^2 = 0.9782 which is very satisfactory. This case shows that not all data can be satisfactorily checked; perhaps Mozart's Sonata had to be partitioned in three parts and all analyzed separately.

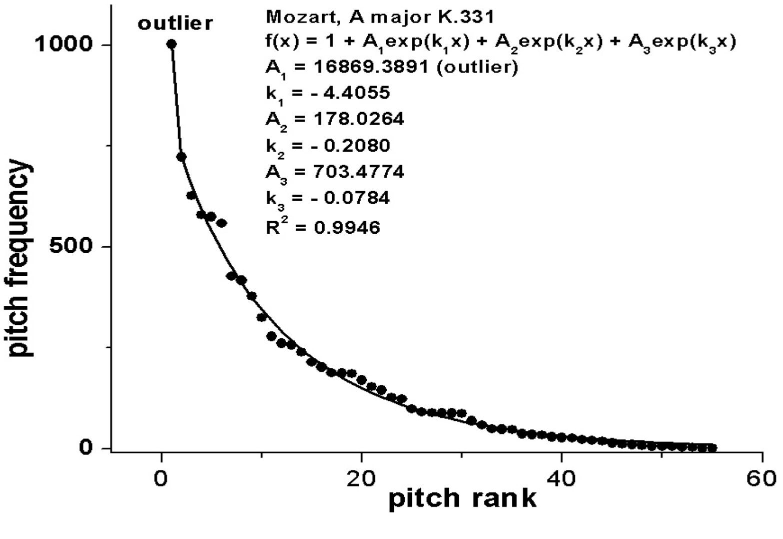

Figure 4. Fitting the pitch rank-frequencies in Mozart's *Sonata A major K.331* with a function of type (6) indicates three strata

Summary

Since Zipf's power function or the corresponding zeta distribution do not always capture satisfactorily the sequence of ranked frequencies, a more satisfactory solution is a sum of exponential expressions which at the same time gives information about the number of strata in the frequencies. The aim of this article was to show that the background linguistic hypothesis concerning changes in the strategy of text creation leads to a differential equation of n-th order. Usually a third order is sufficient but in many cases the fitting itself shows that the order can be reduced. If a text is monolithic, it contains only one stratum. Unfortunately, there are so many aspects of human artefacts - and their number increases with the progress of science - that an enormous number of analyses will be necessary in order to get a more solid basis in this research.

Stratification is, as a matter of fact, a special aspect of self-organization. If something evolves, it gets more complex. Languages and texts are no exceptions. In systems theoretical view, strata are sometimes subsystems evolving in the neighbourhood of and interdependence with other subsystems. For language it is a known fact but for texts it is not that evident because text is a ready product. However, text represents at least two entities: the entity created by the author and the entity interpreted by the reader. The second entity differs with every reader. It is not identical with the written entity - otherwise no "literary science" would exist - and it may change even with one reader. The interpreted text gets part of the mind of the reader and evolves as his mind evolves.

Stratification in language and text has some intersections with diversification, one of the Zipfian forces (cf. Köhler 2005). Everything diversifies in language; the language community and the hearer slow this process down, otherwise the communication would break down. But diversified entities create dialects, sociolects, idiolects, new languages, different presentations of stage plays, new vistas of texts, etc. As a matter of fact, the present article shows merely the stratification process but does not identify the strata.

Acknowledgments. The authors are most grateful to Professor Cristian Calude for his valuable review of this article.

References

Beliankou A., Köhler R. (2010). The distribution of parts-of-speach in Russian texts. *Glottometrics 20, 59-69.*

Fan, F., Altmann G. (2010). On meaning diversification in English. In: *Sprachlehrforschung: Theorie und Empirie. Festschrift für Rüdiger Grotjahn*: *223-233*. Frankfurt: Lang.

Hřebíček, L. (1997). *Lectures on text theory*. Prague: Oriental Institute**.**

Kelih, E. (2009). Graphemhäufigkeiten in slawischen Sprachen: Stetige Modelle, *Glottometrics 18, 52-68.*

Knight, R. (2013). Laws Governing Rank Frequency and Stratification in English Texts. *Glottometrics 25 (present issue).*

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 760-774*. Berlin: de Gruyter.

Köhler, R. (2006). The frequency distribution of the lengths of length sequences. In: Genzor, J., Bucková, M. (eds.), *Favete linguis. Studies in honour of Victor Krupa: 145-152*. Bratislava: Slovak Academic Press.

Köhler, R. (2008). *Word length in text. A study in the syntagmatic dimension*. In: Mislovičová, S. (ed.), *Jazyk a jazykoveda v pohybe*: *416-421*. Bratislava: VEDA: Vydavatel'-stvo SAV.

Köhler, R. (2008). Sequences of linguistic quantities. Report on a new unit of investigation. *Glottotheory 1(1), 115-119*.

Laufer, J., Nemcová E. (2009). Diversifikation deutscher morphologischer Klassen in SMS. *Glottometrics 18, 13-25*.

Martináková, Z., Popescu, I.-I., Mačutek, J., Altmann, G. (2008). Some problems of musical texts. *Glotometrics 16, 80-110*

Nemcová, E., Popescu, I.-I., Altmann, G. (2001). Word associations in French. In: Berndt, A., Böcker, J. (eds.), *Sprachlehrforschung: Theorie und Empirie: 223-237*. Frankfurt: Lang.

Popescu, I.-I., Altmann, G., Köhler, R. (2010). Zipf's law — another view. *Quality and Quantity 44(4), 713-731*.

Popescu, I.-I., Čech, R., Altmann, G. (2011). On stratification in poetry. *Glottometrics 21, 54-59*.

Popescu, I.-I., Mačutek, J., Altmann, G. (2009). *Aspects of word frequencies,* Lüdenscheid: RAM.

Popescu, I.-I., Martináková-Rendeková Z., Altmann G. (2012). Stratification in musical texts based on rank-frequency distribution of tone pitches, *Glottometrics 24, 25-40*.

Sanada H., Altmann G. (2009). Diversification of postpositions in Japanese, *Glottometrics 19, 70-79*.

Tuzzi, A., Popescu, I.-I., Altmann, G. (2010). *Quantitative analysis of Italian texts,* Lüdenscheid: RAM.