Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 95 (2013) 447 – 454

Procedia
Social and Behavioral Sciences

5th International Conference on Corpus Linguistics (CILC2013)

Lexical Statistics and Tipological Structures: A Measure of Lexical Richness

Joan Torruella[a], Ramon Capsada[b]*

[a]*ICREA-UAB, Dp Filologia Espanyola, Universitat Autònoma de Barcelona, Bellaterra 08193, Spain*
[b]*IES Sabadell, C/ Juvenal 2, Sabadell 08206, Spain*

Abstract

For some time now research has been carried out in the field of lexicometry into the statistical indices that enable lexical richness to be evaluated. The main problem lies in the fact that there should be no influence at all in the results of the formula of the length of the text in terms of the number of words it contains. Therefore, different indices have been designed, which are increasingly complex and sophisticated. This work is a review of the most important indices for calculating lexical richness, in order of complexity, looking into whether or not they are dependent on text length and a comparative analysis of the results of the different indices for different text types is presented.

Keywords: lexicometry; lexical richness; text types.

1. Presentation

The appearance, at the end of the last century, of the discipline of corpus linguistics brought new dimensions to the study of language in general and vocabulary in particular, and at the same time made it possible to study the quantitative aspects of texts with a level of precision that had previously been virtually impossible. One of the biggest benefits of this new discipline is lexicometry since it is an applied branch of lexicography that consists in the use of vocabulary according to its quantification. Analysis of lexical quantities and their proportions within texts are a good example of the use of this discipline and of the possibility of researching it using an empirical approach.

* Corresponding author. Tel.: 34+93-581-2962; fax: 34+93-581-1686.
E-mail address: joan.torruella@uab.cat

doi:10.1016/j.sbspro.2013.10.668

But the question lies in working out how to place a value on the lexical quantities and how to find out what the level of profusion of vocabulary is in a given work.

Over the last 100 years different studies have been carried out (Herdan 1960; Guiraud 1960; Carroll 1964; Sommers 1966; Muller 1968; Tuldava 1993; Malvern 2000, 2004; Baayen 2001, 2008; McCarthy 2005; Van Gijsel 2005) and there have been several proposals of statistical formulae for calculating the lexical richness of texts, all of them aimed at avoiding the problem that their length could be a conditioning factor in the results and that texts of different lengths could be compared.

First, this article presents the most important indices used for evaluating lexical richness and then offers some of the experiments that have been carried out in the area of testing for lexical richness for different text types in texts taken from a corpus divided into three categories, one of which is typological. It is the *Corpus Informatitzat del Català Antic* (CICA) (Computerised Corpus of Old Catalan). The idea is to test whether the text type influences the value of lexical richness in the texts.

We believe that the results can offer a new parameter that contributes to obtaining data that can help in the cataloguing of texts in corpora.

2. Different indices for calculating lexical richness in texts

In the last sixty years there have been a series of calculation proposals for measuring the lexical richness of a text. This richness gives us an idea of the number of different terms used in a text and the diversity of the vocabulary.

There is a **first class** of indices based on the direct relationship between the number of terms and words (*type-token*).

TTR (*type-token ratio*) (1957, Templin)

$$\mathrm{TTR} = \frac{t}{n}$$

Then the **TTR** formula underwent different simple corrections:

RTTR (*root type-token ratio*) (1960, Giraud)

$$\mathrm{RTTR} = \frac{\mathrm{t}}{\sqrt{\mathrm{n}}}$$

CTTR (*corrected type-token ratio*) (1964, Carrol)

$$\mathrm{CTTR} = \frac{\mathrm{t}}{\sqrt{2\mathrm{n}}}$$

More recently a **second class** of indices has been developed using formulae based on logarithmic function. This function grows in such a way as to adapt better to the behaviour of the relation that exists between the terms (*types)* and the total number of words in a text (*tokens*).

(1960) Herdan : $\mathrm{H} = \frac{log\ t}{log\ n}$ (1966) Summer: $\mathrm{S} = \frac{\log(log\ t)}{\log(log\ n)}$

(1966) Mass : $\mathrm{M} = \frac{log\ n - log\ t}{log^2\ n}$ (1978) Dugast : $\mathrm{U} = \frac{log^2 n}{log\ n - log\ t}$

(1993) Tuldava: $\mathrm{T} = \frac{log(log\ t)}{(log(log\ \frac{n}{t} + A))^5}$ ***A*** is a parameter that depends on the genre.

Of these five indices, the one that displays most stability with respect to the text length is that of **Mass**.

A **third class** of indices is formed by a group of indices obtained from more complex calculations.

MSTTR (mean segmental type-token ratio) (1944, Johnson).

In this process the text to be analysed is divided into equal segments in terms of the number of words (normally 100 words per segment). For each segment the TTR is calculated and using an arithmetic mean of the TTR for each segment the MSTTR is obtained.

MTLD (measure of textual lexical diversity) (2005, McCarthy).

The starting point for this index is similar to that of the MSTTR, since the text is also divided into segments and the TTR is calculated for each; but in this case the length of the text is variable and depends precisely on the value that the TTR is displaying as the segments are extended. Each segment ends when its TTR reaches a value of 0.72.

At the end of the text the $MTLD = \frac{L}{n}$, calculation is applied, where L is text length in number of words and n is the number of segments.

HD-D.

To calculate this index small parts of the text are always used, calculating their average TTR, but unlike the MSTTR and the MTLD it does not use sequential segments but samples made up of words selected at random, and therefore from all over the text. For technical reasons the length is set at 42 words which can be taken from anywhere in the text. The HD-D index is therefore the average TTR of all of these. Given the huge number of possible samples, the average is not calculated directly but via the calculation of probabilities using the hypergeometric probability distribution.

3. Indices that depend on text length

The biggest problem when studying lexical richness of texts is finding calculation methods that do not depend on the length of the text.

In this work, the first step was to see if the resulting value of each of the indices presented depended on the text length or not.

To do that a work of 418,301 words was elected and divided into 17 blocks of 24,606 words each. Then, in order to compare the results, the seven formulae described above were applied.

Table 1. Equal blocks

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	Total
Tokens	24606	24606	24606	24605	24606	24606	24606	24606	24606	24606	24606	24606	24606	24606	24606	24606	24606	418301
Types	3480	3402	3317	3232	3408	3591	3455	3509	3518	3329	3430	3471	3504	3716	3572	3215	3536	16960
	7,070	7,232	7,418	7,612	7,220	6,852	7,121	7,012	6,994	7,391	7,173	7,089	7,022	6,621	6,888	7,653	6,958	24,663
TTR	0.14	0.14	0.13	0.13	0.14	0.15	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.15	0.15	0.13	0.14	0.04
RTTR	22.18	21.69	21.15	20.60	21.73	22.89	22.03	22.37	22.43	21.22	21.87	22.13	22.34	23.69	22.77	20.50	22.54	26.22
CTTR	15.69	15.34	14.95	14.57	15.36	16.19	15.57	15.82	15.86	15.01	15.46	15.65	15.80	16.75	16.10	14.49	15.94	18.54
Mass	0.019	0.019	0.020	0.020	0.019	0.019	0.019	0.019	0.019	0.020	0.019	0.019	0.019	0.018	0.019	0.020	0.019	0.019
MSTTR	0.70	0.70	0.69	0.69	0.70	0.71	0.71	0.71	0.71	0.71	0.72	0.72	0.71	0.72	0.68	0.69	0.69	0.70
MTLD	82.246	80.467	75.268	75.721	77.061	83.497	86.292	85.003	84.958	90.836	96.165	92.259	86.378	92.737	72.920	73.836	70.510	82.057
HD-D	0.827	0.829	0.823	0.816	0.820	0.834	0.830	0.831	0.832	0.831	0.833	0.836	0.830	0.833	0.817	0.816	0.815	0.830

The table shows that all the formulae show similar numbers for each block but in the first three cases (TTR, RTTR and CTTR), the value changes substantially when the calculation is made for the whole work (last column). However, in the other four cases (Mass, MSTTR, MTLD and HD-D), the value for the total stays within the maximum and minimum limits for the different blocks.

The difference between the indices for which the total value is different from the partial values and those for which it is not, or rather those that are text-length sensitive (TTR, RTTR and CTTR) and those that are not (Mass,

MSTTR, MTLD and HD-D), can be seen more clearly when the work is divided into 17 cumulative blocks until it has all been covered.

Table 2. Cumulative blocks

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Total
Tokens	24606	49212	7388	98423	123029	147635	172241	196847	221453	246059	270665	295271	319877	344483	369089	393695	418301
Types	3480	5266	6529	7578	8615	9749	10556	11412	12196	12806	13468	14049	14675	15312	15974	16404	16960
TTR	0.14	0.11	0.09	0.08	0.07	0.07	0.06	0.06	0.06	0.05	0.05	0.05	0.05	0.04	0.04	0.04	0.04
RTTR	22.18	23.74	24.03	24.15	24.56	25.37	25.43	25.72	25.92	25.82	25.89	25.85	25.95	26.09	26.29	26.14	26.22
CTTR	15.69	16.79	16.99	17.08	17.37	17.94	17.99	18.99	18.33	18.25	18.31	18.28	18.35	18.45	18.59	18.49	18.54
Mass	0.019	0.019	0.019	0.019	0.019	0.019	0.019	0.019	0.019	0.020	0.019	0.019	0.019	0.019	0.019	0.019	0.019
MSTTR	0.70	0.70	0.70	0.69	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.71	0.70	0.70	0.70
MTLD	82.246	81.391	79.154	78.293	78.083	79.028	79.958	80.542	80.975	81.866	83.009	83.715	83.917	84.494	83.600	82.908	82.057
HD-D	0.827	0.829	0.828	0.826	0.825	0.827	0.828	0.829	0.829	0.830	0.831	0.831	0.831	0.832	0.831	0.830	0.830

The results of the first three formulae clearly depend on the length of the text. The result increases or decreases as the text grows.

The last four formulae do not depend on text length. The number fluctuates between two values independent of the length of the text.

This experiment was carried out on different works and the results were the same.

4. Comparison of results between different text types

Can the indices whose results do not depend on text length be useful for analysing lexical richness between different text types or for observing whether there are differences between different authors?

Below are tables showing the comparison between works belonging to the same text type and others that are typologically different.

It should be remembered that as the index value increases the lexical richness also increases, except in the case of Mass where the reverse is true.

Table 3. Fiction

Obra	Curial e Güelfa	Decamerón (1 parte)	Tirant lo Blanch	Llibre de Meravelles
Autor	---	G. Bocaccio	J. Martorell	R. Llull
Tokens	146199	164158	418302	167.009
Types	9574	12418	16961	8.583
Mass	0.019	0.018	0.019	0.021
MSTTR	0.71	0.71	0.70	0.61
MTLD	91.287	87.567	82.057	49.152
HD-D	0.840	0.838	0.830	0.818

Table 4. Chronicles and history texts

Obra	Cròniques d'Espanya	Crònica [B. Desclot]	Llibre dels fets del rei en Jaume	Crònica [R. Muntaner]
Autor	P. M. Carbonell	B. Desclot	Jaume I	R. Muntaner
Tokens	44832	84329	143755	221424
Types	5722	5889	8704	10120
Mass	0.018	0.021	0.020	0.020
MSTTR	0.68	0.67	0.66	0.67

MTLD	68.960	63.843	59.574	65.697
HD-D	0.819	0.816	0.796	0.821

Table 5. Religious and moral works

Obra	Homilies de Tortosa	Les edats i l'epístola de Jhs	Sermons I + II	Vida de santa Caterina	Spill de la vida religiosa	Vita Christi [I. de Villena]	Vides de Sants Rosselloneses	Disputació dels cinc savis
Autor	---	---	V.Ferrer	M. Péreç	---	I de Villena	Anònim	R. Llull
Tokens	504	57111	155764	22072	46709	91378	162795	35806
Types	223	5903	10741	3654	4629	7515	12872	2215
Mass	0.021	0.019	0.019	0.018	0.020	0.019	0.018	0.025
MSTTR	0.68	0.67	0.68	0.74	0.67	0.71	0.67	0.57
MTLD	64.454	65.926	69.252	115.924	64.815	92.228	66.826	39.616
HD-D	0.824	0.827	0.838	0.834	0.844	0.829	0.840	0.786

Table 6. Royal Court prose

Obra	Capítols de greuges per la ciutat d'Oriola	La reintegració de la Corona de Mallorca a la Corona d'Aragó 5	Documents de la Cancelleria d'Alfons el Magnànim	Documents de la Cancelleria d'Alfons III
Autor	Martorell (escrivà)	VVAA	---	---
Tokens	5866	20576	23417	46527
Types	1071	3089	3567	4591
Mass	0.023	0.019	0.019	0.020
MSTTR	0.62	0.69	0.68	0.67
MTLD	54.657	75.761	75.567	69.644
HD-D	0.769	0.819	0.820	0.819

Table 7. Administrative texts

Obra	Els quatre llibres de la reina Elionor	La Germania	El sínode del bisbe Baccallar	Liber Consiliorum
Autor	G. Oliver	VVAA	A. Baccallar	---
Tokens	22713	33944	36488	45649
Types	1182	4332	4863	3563
Mass	0.029	0.019	0.018	0.022
MSTTR	0.63	0.68	0.70	0.67
MTLD	54.104	70.050	80.410	65.054
HD-D	0.742	0.823	0.824	0.826

Table 8. Legal texts

Obra	Usatges de Barcelona	Llibre del Consolat de Mar	Furs de València	Costums de Tortosa	Ordinacions con los reys e reynas d'Aragó se consagren	Ordinacions de la Casa i Cort de Pere el Cerimoniós
Autor	---	---	Jaume I	---	---	---
Tokens	9561	56352	130086	166735	7529	74280
Types	1604	3179	6744	8305	1263	6407
Mass	0.021	0.024	0.021	0.021	0.022	0.019
MSTTR	0.67	0.60	0.62	0.62	0.63	0.69
MTLD	64.471	50.679	49.728	51.432	51.578	74.828
HD-D	0.837	0.801	0.829	0.827	0.789	0.837

Table 9. Court texts

Obra	Llibre de Cort de Justícia de Cocentaina 1	Llibre de Cort de Justícia de Cocentaina 2	Clams i crims a la València medieval 1	Clams i crims a la València medieval 2	Llibre d'Inquisi-cions de Castellitx	Llibre de Cort de Justícia d'Alcoi (	Llibre de Cort de Justícia de València
Autor	VVAA	---	VVAA	VVAA	---	VVAA	VVAA
Tokens	58740	117806	48146	45584	10146	18392	29371
Types	4274	6075	4581	4013	1380	2341	2677
Mass	0.022	0.022	0.020	0.021	0.023	0.021	0.023
MSTTR	0.64	0.63	0.64	0.65	0.64	0.63	0.63
MTLD	56.794	57.821	58.902	60.774	59.530	49.991	56.847
HD-D	0.801	0.803	0.808	0.807	0.806	0.799	0.796

Table 10. Scientific and technical texts

Obra	Aforismes 2	Libre de coch	Quesits o perquens	Cànon d'Avicenna	Receptari	Llibre de confits	Tractat d'astrologia	Començaments de medicina
Autor	Hipòcrates	M. Robert	G. Manfredi	---	J.Martina	---	B. Tresbéns	R. Llull
Tokens	8079	41529	82148	128016	4441	6795	906	32117
Types	1595	2611	5614	7635	1045	988	401	2260
Mass	0.020	0.024	0.021	0.020	0.021	0.025	0.018	0.025
MSTTR	0.64	0.64	0.60	0.62	0.64	0.64	0.66	0.56
MTLD	50.842	53.351	50.221	45.187	50.993	54.464	54.278	37.289
HD-D	0.818	0.784	0.799	0.788	0.773	0.801	0.776	0.773

Table 11. Epistolary works

Obra	Carta del cavaller Llàtzer	Carta d'Arnau d'Erill 1	Carta de Berenguer Batle	Carta d'Arnau d'Erill 2	Cartes al Bisbe d'Urgell	Epistolari de Ferran I d'Antequera	Cartes dels Borja	Epistolaris d'Hipòlita de Liori
Autor	Lloscos	A. Erill	Batle	A. Erill	---	VVAA	VVAA	VVAA
Tokens	320	378	467	468	506	136085	38408	145913
Types	187	188	176	216	227	9076	4522	8778
Mass	0.016	0.020	0.026	0.020	0.021	0.019	0.019	0.020
MSTTR	0.74	0.67	0.67	0.67	0.67	0.70	0.70	0.71
MTLD	100.348	68.393	62.986	75.896	62.109	82.977	85.007	91.351
HD-D	0.823	0.804	0.773	0.807	0.783	0.826	0.838	0.823

Table 12. Poetry

Obra	Trobes en lahors de la V. M.	Lo passi	Spill	Poesies	Oració	Contemplació	Lo procés de les olives	Breu descripció
Autor	---	Fenollar, y P. Martines	J. Roig	A. March	J. R. Corella,	B. Fenollar i J. Escrivà	Fenollar, *et alt.*	Montmajor
Tokens	16521	30785	45071	77283	527	3400	3541	2167
Types	3436	4621	10851	7625	303	1129	1145	944
Mass	0.017	0.018	0.012	0.018	0.014	0.017	0.017	0.014
MSTTR	0.78	0.76	0.83	0.75	0.77	0.79	0.76	0.78
MTLD	144.631	125.463	265.419	118.830	155.109	157.282	129.783	165.653
HD-D	0.874	0.878	0.909	0.866	0.871	0.870	0.859	0.866

An analysis of the results of lexical richness seen in the tables above shows that, in general, the different indices give very similar results for each of the different text types analysed, albeit on different scales. Except in the odd

case, the maximum and minimum richness values coincide in the four indices and where they do not the difference is of no great significance.

As far as the values of lexical richness between different text types is concerned, it is observed that is , for some types of text, there are no significant differences (for example between Royal Court prose and legal texts) in other there are (for example poetry is by far and away the richest and scientific prose the poorest). These typological differences can be intuited beforehand, but it is important to see that there are now indices available that are capable of detecting and quantifying them.

In the data provided in the previous tables it can also be observed that when an author presents a low index of richness this is independent of the type of text they are writing. This is the case of Ramon Llull who displayed a low index of richness in all the texts types in which his works have been analysed (fiction prose, religious works and scientific texts).

5. Conclusions

After revising some of the most relevant indices for calculating lexical richness and experimenting with texts of different lengths and types, it can be said that of the seven indices used in this work (TTR, RTTR, CTTR, Mass, MSTTR, MTLD y HD-D) the first three are not valid for studying lexical richness of texts since they depend on text length. Of the four that are unaffected by text length, some are more sensitive than others (Mass is shows low sensitivity and MTLD very high sensitivity), understanding sensitivity as the level of detail in the measurement of lexical diversity.

With respect to the four indices that do not depend on the length of the text the maximum and minimum values for each index always are always the same, or at least very similar. The index that departs most from the rest is HD-D, although only in its upper range and by very insignificant values.

The values of the four indices used in the analysis of the lexical richness in different works of different types show that these indices are capable of detecting and quantifying the differences in lexical richness between text types and confirm that if in some cases the values between different text types show slight differences, in other cases they are significant: poetry is the text type with most lexical richness, while the court texts and, above all, all the scientific texts are those displaying the least lexical richness. In terms of individual authors, it is observed that the results are more diverse than for text types. According to the tables, where a lack of lexical richness occurs it is present throughout all the text types written by that author.

We believe that the increase in recent text corpora and the possibilities offered by computer technology offer today's vocabulary scholars and the recent studies of indices for the calculation of text richness and a good opportunity for lexicometrical research. In this work we have attempted to provide a first sketch of the possibilities offered to researchers by theses indices and their effect on in different text types. We are aware that our simple size was small for arriving at definite conclusions, but we believe that it is useful as a starting point in approaching a subject and line of research that we hope will continue.

References

Baayen, R. H (2001). *Word frequency distributions*. Dordrecht [etc.] b: Kluwer Academic Publisher.

Baayen, R. H (2008). Analyzing linguistic data: a practical introduction to statistics using R. Cambridge: Cambridge University.

Carroll, J. B. (1964). Language and Thought. Englewood Cliffs : Prentice-Hall, Inc.

Francais Moderne, 46, 25-32.

Guiraud, P. (1960). Problèmes et Méthodes de la Statistique Linguistique. Paris: Presses universitaires de France.

Herdan, G. (1960). Quantitative Linguistics. London: Butterworth.

Malvern, D., et alt. (2000). Measuring Vocabulary Diversity Using Dedicated Software. Literary and Linguistic Computing; 15, 3, 323-337.

Malvern, D., et alt. (2004). Lexical Diversity and Language Development. Quantification and Assessment. New York: Palgrave Macmillan.

McCarthy, P. M. (2005). An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD). Dissertation Abstracts International, 66 12, (UMI No. 3199485).

Muller, C. (1968): Initiation à la statistique linguistique. Paris: Librairie Larousse.

Somers, H. H. (1966). Statistical methods in literary analysis. In J. Leeds (Ed.), The computer and literary style, pp. 128-140. Kent, OH: Kent State University.

Torruella, J, Pérez Saldanya, M. & Martines, J. (Dirs.) (2013). CICA = Corpus Informatitzat del Català Antic. <http://www.cica.cat>.

Tuldava, J. (1993). The statistical structure of a text and its readability. In L. Hrebicek & G. Altmann (Eds.), Quantitative text analysis, pp. 215-227. Trier: Wissenschaftlicher Verlag Trier.

Van Gijsel, Sofie, et alt. (2005). A variationist, corpus linguistic analysis of lexical richness. Quantitative Lexicology and Variational Linguistics. <http://www.bhamlive2.bham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2005-journal/TheLexicon/LexicalRichness.doc>.