Can Type–Token Ratio be Used to Show Morphological Complexity of Languages?

Article *in* Journal of Quantitative Linguistics · June 2014

DOI: 10.1080/09296174.2014.911506

1 author:

Kimmo Kettunen

94 PUBLICATIONS **706** CITATIONS

SEE PROFILE

Can type-token ratio be used to show morphological complexity of languages?

Kimmo Kettunen [1]

Abstract

Type-token ratio (TTR), also known as vocabulary size divided by text length (V/N) is a simple measure of lexical diversity. It has been used in literary studies, studies in child language and even psychiatry (Covington and McFall, 2010). The basic problem of TTR is that it is affected by the length of the text sample. Several suggestions for improving this fault have been given, including standardizing the length of text samples, using logarithms in the basic formula, etc. (cf. Covington and McFall, 2010; Tweedie and Baayen, 1998).

We show in this paper that simple TTR and its more elaborate calculation MATTR (Covington and McFall, 2008, 2010) can be used for approximation of morphological complexity of languages. This usage of TTR has been notified in Juola (1998, 2008) with analysis of six languages. We analyze text material with TTR and MATTR from two differing sources: firstly, text of the EU constitution with 21 languages (Sadeniemi et al., 2008) and secondly with 16 of the same languages with available non-parallel random data from the Leipzig corpus (Quasthoff et al., 2006). We compare the automatic analysis results to two independent linguistic measures of morphological complexity. Firstly, we use number of non-homographic noun forms in a language's inflectional paradigms, the paradigm size (Plank, 1986). Secondly we use available inflectional synthesis figures of verbs produced by the AUTOTYP project (Bickel and Nichols, 2004, 2005).

We enrich our corpus findings with data from information retrieval (IR) results. McNamee et al. (2009) and Kettunen (2008, 2013) have suggested that improvements in achieved IR effectiveness with usage of word form variation management depend on the morphological complexity of the languages. Thus this IR gain data can be used to give independent evidence to evaluation of morphological complexity.

Our results show that earlier Juola complexity figures and TTR and MATTR calculations correlate moderately in the EU constitution data. Figures given by TTR and MATTR correlate highly with each other in both corpora, and they also correlate highly with the number of non-homographic noun forms in a language. Correlation to inflectional synthesis of the verbs was found weakly positive in most cases, but the data was scarce. All the three computed measures are able to order the languages quite meaningfully in a morphological complexity order that at least groups most of the languages with same kind of languages and the most and least complex languages are clearly separated. It seems also that TTR and MATTR order the languages quite consistently with both corpora. In the conclusion we discuss how the complexity figures can be utilized.

Introduction

Linguistic complexity and its specific sub-parts, such as morphological, syntactic, morpho-syntactic, typological etc. complexity, has gained much interest during the last ten years. Different views of linguistic complexity have been proposed and also different ways of approximating the complexity of different languages or their subsystems have been suggested. As a whole, the notion of linguistic complexity is not easily definable or measurable, but some of its subparts offer more clearly definable and operationalizable targets for analysis.

[1] +358 50 5710859 Kiviteltankatu 6 A 11 FIN-00710 Helsinki kkettun4@welho.com

On morphological level linguistic complexity means roughly, that the language has lots of inflection, which is realized, for example, in the number of different nominal case forms the language has (data for different languages for example in Iggesen, 2011). Or put in another way, more structural units, rules or representations mean more complexity (Hawkins, 2009). This is in accordance with a general definition of complexity as number and variety of elements (Sinnemäki, 2012: 16). Finnish, for example, has 14 different morphological cases, Hungarian 19-21 and English has two. This means that Finnish and Hungarian have many variating noun forms, as English has only a few due to the nature of the case morphology. The number of different possible forms for a basic inflected noun – without clitics or possessive endings - of the EU languages varies from two (Dutch, Spanish and other Romance languages) to about 38-42 (Hungarian). Many times already the number of cases in the language is indicative of morphological complexity, but not always (e.g. in the case of Swedish, Danish and Bulgarian). Then other morphological categories, such as marking of definiteness and expression of number in the language, are the key factors (Stump 2001). Compounding, creation of new words by concatenating existing words to form new ones, gives added complexity to some languages, e.g. for Swedish, Dutch and German. When we talk about morphological complexity of a language, we are talking about local complexity, complexity of some part of an entity. Study of local complexity of languages is considered to be more feasible than study of global complexity, which can be considered methodologically and practically unattainable (Sinnemäki, 2012).

There have been a number of suggestions how to measure morphological complexity of languages. However, no definition or quantification of the complexity of a linguistic system or sub-system is widely accepted. Here we present only a few suggestions that have an algorithmic implementation.

A popular algorithmic way to approximate the morphological complexity of a language has been Patrick Juola's (1998, 2008) suggestion of distorting word structures by using a unique random number for each different word type. After distortion, the data is compressed using a compression algorithm. Then the size of the compressed original word file is divided by the size of the compressed distorted word data file. The result tells the complexity of each language's morphology on the basis of Kolmogorov complexity that the compression algorithm approximates. Besides by Juola, the method has been used, for example, by Sadeniemi et al. (2008) and Ehret and Szmrecsanyi (to appear[2]) for a set of languages with mainly plausible results. Ehret and Szmrecsanyi also show that the Juola method suits equally well to parallel, semi-parallel and non-parallel texts, the last being a good addition to the method's applicability. Besides Ehret and Szmrecsanyi show that the method is robust with respect to different sampling points in the process.

Bane (2008) suggests another kind of approach for approximation of morphological complexity of languages. He uses Linguistica, software that induces morphology of a language from a given text sample. Based on the analysis of an un-annotated text

[2] Strictly speaking Ehret and Szmrecsanyi use a variant of the Juola method, where 10 % of the characters in the samples are removed randomly and the resulting file is compressed. This approach is introduced in Juola (2008) and evidently it is comparable to the distortion method of Juola (1998).

corpus, Linguistica separates word stems, affixes and signatures. In this approach the morphological complexity of the language is based on the formula

$$\textbf{\textit{Morphological complexity}} = (DL(Affixes)+DL(Signatures)) / (DL(Affixes)+DL(Signatures)+DL(Stems))$$

Affixes and stems in the formula describe linguistic units identified by the Linguistica software, signatures describe the "*possible distributions of affixes upon stems*". Thus the method measures "*a language's morphological complexity as the proportion of the lexicon's total description length"*, DL in the formula, "*that is due to the description lengths of the affixes and signatures.*" (Bane 2008). Bane's kind of analysis of morphological complexity can be seen as more linguistically oriented than that of Juola's. The results Bane shows for 20 Bible translations seem otherwise plausible, but Romance languages seem to get quite high values.

Patrick Juola (1998, 2008) has notified that morphological complexity ordering of language samples given by the complexity-theoretic analysis are the same as produced by the number of word types in the same samples (and thus also identical reversed with the ordering produced by the number of word tokens). Juola's sample (1998) consists of six languages (Maori, English, Dutch, French, Russian, and Finnish) and his data is from Bible translations of these languages. The suggestion is interesting, but as the data in Juola is smallish, more evidence for evaluation is needed. Bane (2008) has shown with an analysis of 14 Bible translations that TTR order relates also to his complexity measure, although the order of languages by the type-token and the complexity figure is not exactly the same in a bigger sample.

We show in this paper that simple TTR and its more elaborate calculation MATTR (Covington and McFall, 2008, 2010) can be used for approximation of morphological complexity of languages. We analyze text material with TTR and MATTR from two differing sources: firstly, text of EU constitution with 21 languages (Sadeniemi et al., 2008) and secondly with 16 of the same languages with available non-parallel random data from the Leipzig corpus (Quasthoff et al., 2006). We enrich our corpus calculations with data from IR results. McNamee et al. (2009) and Kettunen (2009, 2013) have suggested that improvements in achieved IR effectiveness with usage of word form variation management depend on the morphological complexity of the languages. Thus this IR gain data can be used to give independent evidence to evaluations of morphological complexity.

The paper is organized as follows. We first describe our data and methods. Results of the paper are shown in chapter three, and in the final chapter we discuss the results and how they could be used in practice.

2. Data and methods

Kettunen et al. (2006) and Sadeniemi et al. (2008) used EU constitution and its translation to 21 EU languages in their complexity analysis with the Juola method. Each translation of the constitution text consists of ca. 113 000 – 177 000 word forms and ca. 9100 – 15 000 sentences depending on the language (Kettunen et. al., 2006, Sadeniemi et al., 2008). We re-use the material and its morphological complexity results in our experiments as such. The material represents a parallel corpus approach for

morphological complexity approximation of the languages within one textual genre, that of legal texts.

To add non-parallel data to our evaluation we sampled as many of the same EU languages as were available from the Leipzig corpus pages (Leipzig Corpora Collection Download Page). We found material for 16 of the EU languages. There would have been material for Maltese, but as the Leipzig Maltese data is in the Arabic script, we did not use it.

The Leipzig corpora materials differ very much from the EU constitution corpus. Leipzig corpora consist of newspaper and web material sentences that are not related to each other, and thus they are not proper texts. "*The corpora are identical in format and similar in size and content. They contain randomly selected sentences in the language of the corpus and are available in sizes of 100,000 sentences, 300,000 sentences, 1 million sentences etc. The sources are either newspaper texts or texts randomly collected from the web. The texts are split into sentences. Non-sentences and foreign language material was removed.*" (Quasthoff et al., 2006). We used 10 000 sentence samples, which have about 115 000 to 209 000 words, depending on the language. For some of the languages we used web texts, for some newspaper text, depending on the availability. We'll discuss the possible effects of this kind of random sentence material later.

Our basic data, number of words and word types for each language in both our corpora together with non-homographic noun forms in the language and inflectional synthesis figures are listed in Table 1. Token figures were given by the MATTR software; types were counted from sorted word files without duplicate forms.

	Tokens, EU constitution	Types, EU constitution	Tokens, Leipzig corpus	Types, Leipzig corpus	Number of distinct noun forms in the language's inflectional paradigm	Inflectional synthesis (IS) score for the language
CS	113377	14553	141137	44586	10	N/A
DA	133266	10572	164395	30673	8	N/A
DE	132794	11608	164776	35255	4	2
EL	149030	11618	N/A	N/A	5	5
EN	153703	8273	188140	23573	4	2
ES	163613	9613	172106	23731	2	4
ET	101185	15469	133831	39293	28	N/A
FI	101264	15976	115775	45923	26	3

FR	162589	8935	171097	28126	2	4
GA	158363	10883	N/A	N/A	4	2
HU	123164	14895	N/A	N/A	38	5
IT	149635	9488	186677	30268	2	N/A
LT	109800	14070	142315	41880	12	N/A
LV	112984	12056	137689	35458	9	N/A
MT	157234	13947	N/A	N/A	6	N/A
NL	155004	9453	N/A	N/A	4	N/A
PL	122678	13947	145711	36279	10	N/A
PT	150003	9381	209932	26556	2	N/A
SK	113728	16326	160084	45905	10	N/A
SL	121721	12296	182726	40443	8	N/A
SV	129728	11331	140184	31027	8	N/A

Table 1. Word data

Both sets of our data are roughly of the same size in words, but otherwise totally different in nature. EU constitution is a text from a rather narrow genre of law, texts from the Leipzig corpus are more general but as the sentences are randomly selected, they will give a different type of view. By using these two different materials we wish to show that TTR and MATTR are able to show morphological complexity of a language reliably enough regardless of the used material.

To get a simple comparison measure for each language's morphological complexity we use number of non-homographic noun forms for each language's inflectional paradigm. Nouns are the largest open word form class in any language, and majority of words in texts or dictionaries are nouns. A plausible figure for the number of nouns in texts could be 30-40 %, based on calculations of Hudson (1994) for English, and Kettunen (2005) for Finnish with several tagged corpora for each language. We have gathered the number of noun forms by multiplying case forms of the language with size of the morphological category number in the language (usually two, singular and plural, some times three, as in Maltese and Slovene). For Swedish and Danish we have also included definiteness of the forms in the calculations. Homographic forms in declensions have been omitted from calculations with a check from declension tables. Figures are not exact in many cases, as there are many differences in case inflection for example for different genders, but for our purposes the figures show the right magnitude well. It should also be noted, that as adjectives, numerals and pronouns inflect most of the times in the noun cases, this will increase the occurrence of the different nominal forms in the texts of the languages.

Another linguistic comparison measure is given by the inflectional synthesis of the verbs for eight languages in our corpus, produced by the AUTOTYP project (Bickel and Nichols, 2004, 2005). AUTOTYP's measure of inflectional synthesis shows the degree to which verbs can be marked by inflectional categories. Verb related grammatical categories (e.g. tense, voice and agreement) can be expressed in languages either by individual words or by affixes attached to a word or a stem. Synthetic means that morphemes are combined for this purpose, analytic that the morphemes are left uncombined. English future tense *will show,* for example, is analytic, but past tense *showed* is synthetic. Inflectional synthesis of the verbs gives evidence regarding the number of categories that can be morphologically synthesized on the verb in each language. Verbal inflectional synthesis scores have been earlier used as a measure of morphological complexity at least in Shosted (2006), which is a language typological study.

AUTOTYP inflectional synthesis figures are based on analysis of maximally inflected verb forms, and give the number of inflectional categories that can be attached to the verb, categories per word (cpw). English verbs, for example, show person agreement (he eat***s***) and tense (show***ed***), and thus they have an inflectional synthesis degree of 2 cpw.

Computing of TTR for texts is done as follows: number of word form types (i.e. unique string forms) in each text is divided by the number of running word forms of the tokenized text. MATTR of Covington and McFall (2010) elaborates the basic calculation as follows: "*We cut the Gordian knot by computing and averaging the moving average type–token ratio (MATTR)… We choose a window length (say 500 words) and then compute the TTR for words 1–500, then for words 2–501, then 3–502, and so on to the end of the text. The mean of all these TTRs is a measure of the lexical diversity of the entire text and is not affected by text length nor by any statistical assumptions.*" We used MATTR software, version 2.0, available from http://www.ai.uga.edu/caspr/ with the default window size of 500 words. Effect of the window size will not be further studied in the paper, but it may of importance for different types of studies, as Covington and McFall (2010) suggest.

In what follows we make the following assumptions with respect to the data and methods:

- Figures of the Juola method calculated from the EU constitution are a useful measure of morphological complexity. Although the method is not without problems (cf. Moscoso del Prado, 2011), it has been used with quite plausible results in several publications (cf. Sadeniemi et. al, 2008; Ehret and Szmrecsanyi, to appear).

- Usage of coherent texts (EU constitution) and random sentence data (The Leipzig corpus) can be argued for as follows. Coherent prose does not consist of a string of randomly chosen words and sentences, it has structure, which is usually called discourse structure. Discourse structure affects lexical measures of texts, and some measures are affected more than others (Tweedie and Baayen, 1998). Baayen (1996) shows with two different Dutch corpora that already intra-textual cohesion within paragraphs is sufficient to give rise to substantial deviation between vocabulary sizes in texts with no overall discourse organization. In Baayen's study

a substantial reduction in overestimation of vocabulary sizes is obtained in the Dutch Uit den Boogaart corpus, which contains small text samples, paragraphs taken out of texts, evenly spread over a period of one year. Anyhow, even within paragraphs, words tend to be reused more often than expected. This, in turn, pre-empts usage of other word tokens, among which tokens of types have not been observed among the preceding tokens and leads to a decrease in type richness. In sequences of sentences, words are more likely to be re-used than expected under chance conditions. According to Baayen (1996), "*by randomly sampling individual sentences instead of sequences of sentences, the effects of intra-textual and inter-textual cohesion will largely be eliminated*".

As the Leipzig corpus data consists of randomly chosen sentences, intra-textual or inter-textual cohesion should not affect usage of words. This means that the type-token figures of the Leipzig corpora differ from those of normal texts. There should be more different word types present in the random sentence data than in coherent texts, and thus the Leipzig corpus data will give us a different view. The assumption that randomly sampled sentence data contains more types than coherent texts is confirmed in the Leipzig data. From Table 1 we can see, that in the entire Leipzig corpus data there are about 2.5-3 times more word types in all the languages than in the EU constitution of the same languages. This can be partly affected by the textually restricted style of the EU constitution, but we don't expect the effect of the text genre to be so considerable.

- Parallel and non-parallel data can be used without problems in calculations and comparisons (cf. Ehret and Szmrecsanyi, to appear). In our case we generalize this also to material that is not text, but disconnected sentences taken out of their textual contexts (Leipizig corpus).
- Calculation of noun paradigm size of a language is a plausible simple comparison method for the more elaborate calculations of complexity. Although the paradigm size approach ignores that there are sometimes very regular relations between different inflected forms, this does not matter in our case, and the approach offers a clear numeric value for comparisons. Sizes of the paradigm of different languages vary from minimal (English, French, Spanish) to extra large (Estonian, Finnish and Hungarian), most of the languages belonging to the medium sized group (Plank, 1986).
- Our other linguistic measure of morphological complexity, inflectional synthesis score of verbs produced by the AUTOTYP project, gives us another independent measure of morphological complexity. Verbs also belong to the most frequent word classes in texts, and they give another view of morphological complexity beside noun inflection.
- Corpora sizes used in calculations are roughly the same, but there is some variation in sizes both in parallel and non-parallel data. It should be kept in mind, that the basic TTR calculations are heavily dependent on the size of the corpora (Covington and McFall, 2010; Tweedie and Baayen, 1998). The TTR results shown with our

data may be different with other corpora of different size. Covington and McFall's (2008, 2010) MATTR is supposedly free of this defect.

3. Results

Table 2 shows results of TTR and MATTR calculations for the data. Juola complexity figures in column 2 are from Kettunen et al. (2006). We were not able to compute Juola complexity figures for the Leipzig data.

	Juola complexity EU constitution	TTR EU constitution	MATTR EU constitution	TTR Leipzig	MATTR Leipzig
CS	1.09	0.13	0.57	0.32	0.80
DA	1.13	0.08	0.45	0.19	0.65
DE	1.17	0.09	0.47	0.21	0.70
EL	1.15	0.08	0.47	N/A	N/A
EN	1.05	0.05	0.39	0.13	0.65
ES	1.06	0.06	0.40	0.14	0.61
ET	1.10	0.15	0.59	0.29	0.82
FI	1.16	0.16	0.60	0.4	0.86
FR	1.06	0.05	0.42	0.16	0.66
GA	1.06	0.07	0.45	N/A	N/A
HU	1.14	0.12	0.53	N/A	N/A
IT	1.05	0.06	0.46	0.16	0.68
LT	1.08	0.13	0.59	0.29	0.84
LV	1.08	0.11	0.54	0.26	0.80
MT	1.11	0.09	0.44	N/A	N/A
NL	1.12	0.06	0.41	N/A	N/A
PL	1.14	0.11	0.55	0.25	0.78
PT	1.07	0.06	0.43	0.13	0.65
SK	1.11	0.14	0.59	0.29	0.78
SL	1.06	0.10	0.53	0.22	0.73
SV	1.13	0.09	0.46	0.22	0.68
Mean	**1.10**	**0.10**	**0.49**	**0.23**	**0.73**
Std. deviation	**0.04**	**0.03**	**0.07**	**0.07**	**0.08**

Table 2. TTR, MATTR and Juola complexity figures for the data

TTR, MATTR and Juola complexity figures of the EU constitution correlate moderately in Spearman rank-order correlation test (0.49 and 0.41, respectively, $p < 0.05$). TTR and MATTR calculations correlate highly with each other: 0.97 with the EU constitution figures and 0.93 with the Leipzig corpus figures ($p < 0.001$).

To start with the analysis of the results, let's take a look at orderings of the languages using each measure. Figures 1-3 show the data for 21 languages of the EU constitution. Order of the languages in Figure 1 is by Juola complexity, in Figure 2 by TTR and in Figure 3 by MATTR. Mean numbers for each calculation are given below the figures.

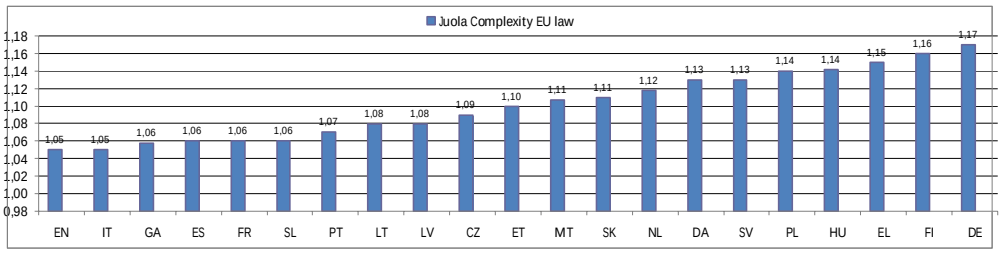

Figure 1. EU constitution complexity figures, order by Juola complexity. Mean 1.10

As there is no gold standard of morphological complexity to make comparisons to, we use the mean figures of calculations and number of noun forms in the language (Figure 8) as a comparison when we discuss the results of each measure.

If we take a look at the Figures 1-3, it seems that at least Danish, Dutch and Swedish are higher in the Juola complexity order than they probably should be, over the mean figure. The same is true of Maltese and especially Greek. In the two other figures all the five are getting lower scores. Slovenian and Estonian seem to be lower than expected in the Juola complexity figure, Slovenian being 0.04 points below the mean and Estonian on the mean.

In Figures 2 and 3 Swedish and Danish are lower and the overall order seems more realistic. Maltese is still quite high, but below the mean, and Greek has a lower position, too. German is in the TTR and MATTR calculations much lower than in the Juola complexity, close to Danish and Swedish, genetically related languages. All the languages that are over the mean or on it in Figures 2 and 3 it seem to be the most complex ones according to the number of nominal forms.

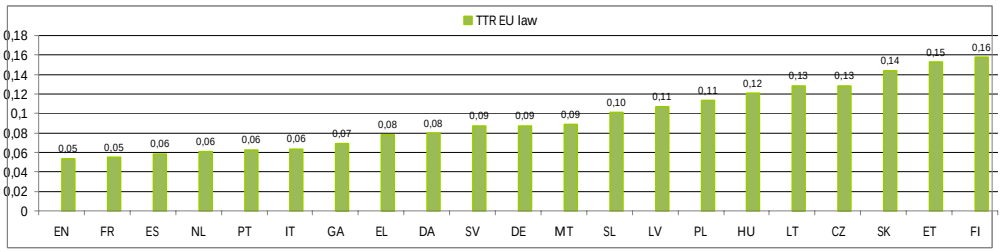

Figure 2. EU constitution complexity figures, order by TTR. Mean 0.10

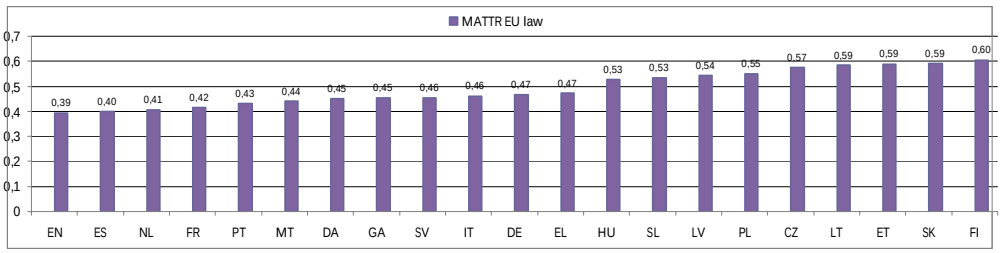

Figure 3. EU constitution complexity figures, order by MATTR. Mean 0.49

Table 3 shows the complexity orders by each measure from least complex to most complex in the EU constitution analysis.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**2**
Juola	EN	IT	GA	FR	ES	SL	PT	LV	LT	CS	ET	MT	SK	NL	DA	SV	PL	HU	EL	FI	D
TTR	EN	FR	ES	NL	PT	IT	GA	EL	DA	SV	DE	MT	SL	LV	PL	HU	LT	CS	SK	ET	F
MATTR	EN	ES	NL	FR	PT	MT	DA	GA	SV	IT	DE	EL	HU	SL	LV	PL	CS	LT	ET	SK	F

Table 3. Complexity orders of languages by each measure

In Figures 4-7 the five languages missing from the Leipzig data have been omitted and here only figures of the TTR and MATTR are shown.

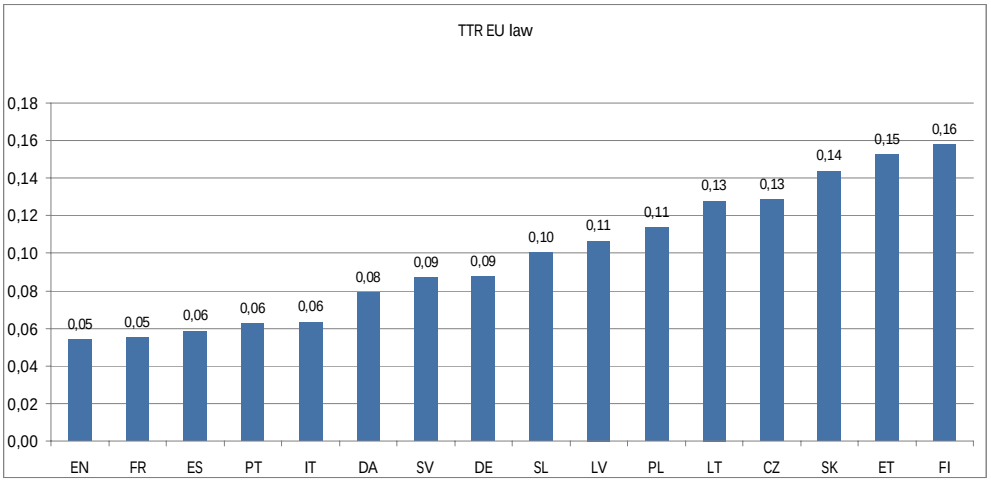

Figure 4. 16 languages of the Leipzig corpora, order by the TTR of the EU constitution. Mean 0.10

In figure 4 the languages are ordered by the TTR results of the EU constitution. The order seems realistic: the most complex languages are on the right side of the mean figure or on it, and Danish, German and Swedish, are near the mean figure. The same happens in Figures 5-7, the only difference being that Italian rises a little in Figure 5. Order of the languages varies a little by each measure.

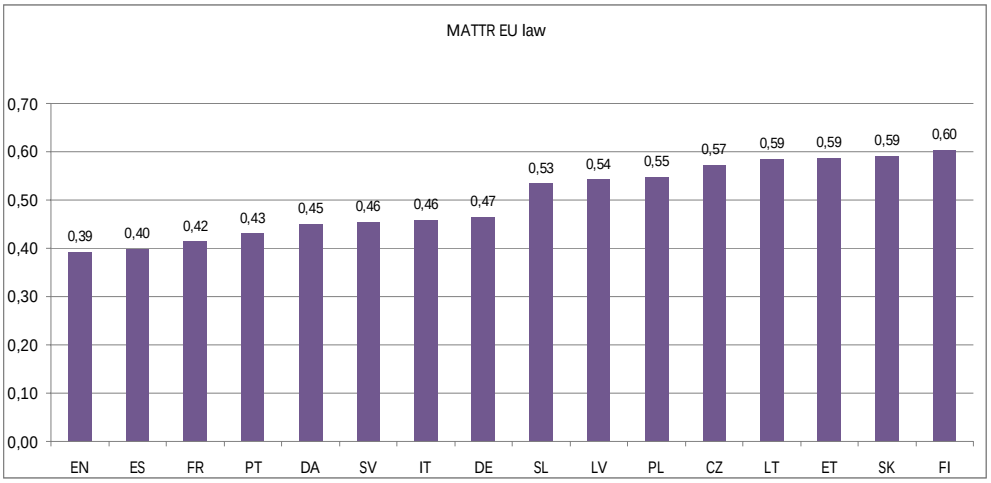

Figure 5. 16 languages of the Leipzig corpora, order by the MATTR of the EU constitution. Mean 0.50.

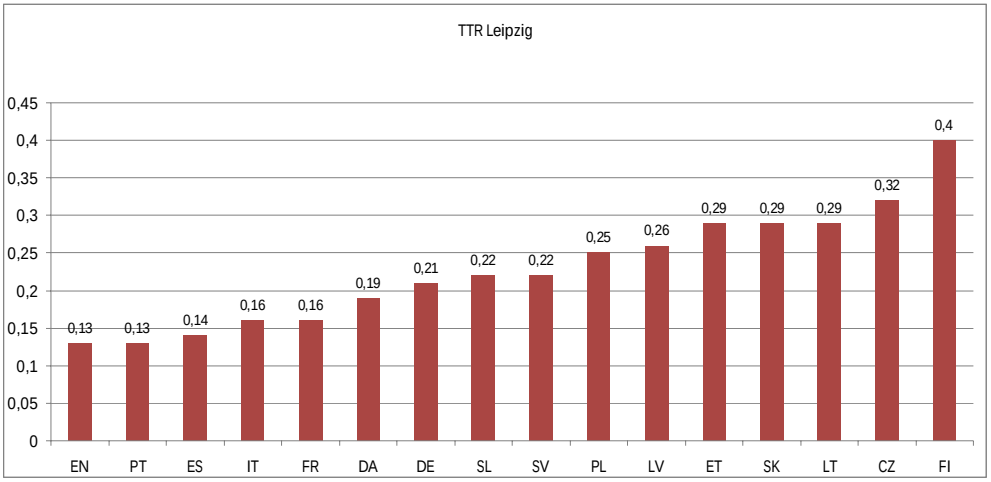

Figure 6. 16 languages of the Leipzig corpora, order by the TTR of the Leipzig corpus. Mean 0.23.

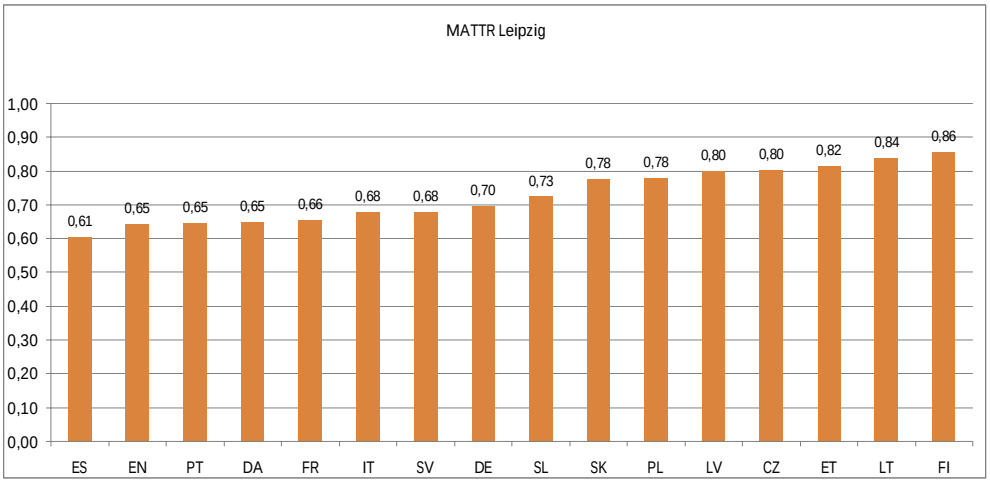

Figure 7. 16 languages of the Leipzig corpora, order by the MATTR of the Leipzig corpus. Mean 0.73.

Finally, in Figure 8, we show the order of languages by the number of distinct noun forms. As can be seen, the languages order here very much in the same way as in the complexity calculations. The mean number of forms is relatively high, due to the three Finno-Ugric languages. Median for the number of forms is 8. Here, again, the languages on the right side of the mean are the most complex, and languages near the mean near the same complexity.

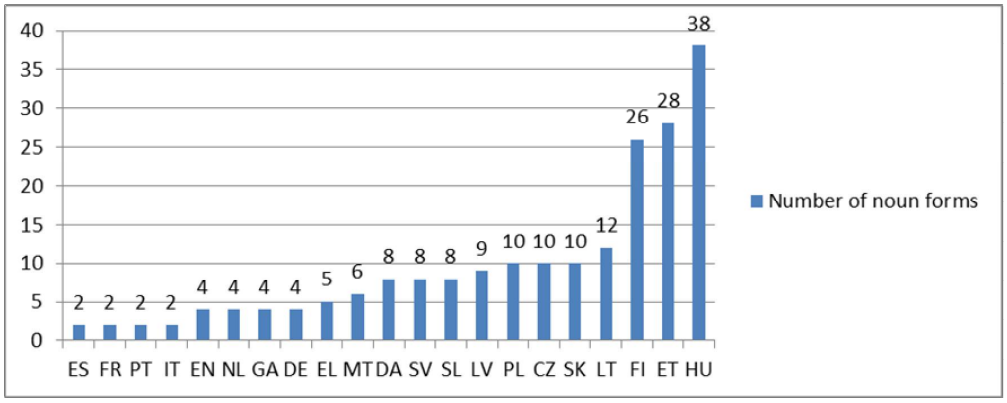

Figure 8. Number of distinct noun forms in the language. Mean figure 9.6

In Table 4 we show Spearman rank-order co-efficients for different complexity calculations vs. number of different noun forms of the language to see how plausible calculations seem in relation to this simple measure.

Correlation between	**Spearman's rank-order correlation**	**Statistical significance**

	co-effcient	
Juola complexity of the EU constitution vs. number of noun forms	0.49	$p < 0.05$
TTR of EU constitution vs. number of noun forms	0.92	$p < 0.001$
MATTR of EU constitution vs. number of noun forms	0.84	$p < 0.001$
TTR of Leipzig corpora vs. number of noun forms	0.91	$p < 0.001$
MATTR of Leipzig corpora vs. number of noun forms	0.88	$p < 0.001$

Table 4. Spearman rank-order correlation co-efficients of complexity figures and number of noun forms

Juola complexity figures from Kettunen et al. (2006) correlate moderately with the number of noun forms. TTR and MATTR figures correlate highly with the number of noun forms in the language in both corpora. Correlations of both TTR and MATTR with the Leipzig corpus material are slightly higher than correlations of the EU constitution.

Spearman rank correlation analysis showed that the verb synthesis figures had a weak but statistically non-significant positive correlation (0.14) with the Juola complexity figure of the EU constitution, TTR of the EU constitution (0.19), and MATTR of the EU constitution (0.3) which all had data for the same 8 languages. TTR and MATTR of the Leipzig corpus lacked data for three languages (EL, GA and HU), and the correlations for these were either non-existent with TTR of Leipzig corpus or weakly negative with MATTR of Leipzig corpus (-0.25). Although the AUTOTYP verb data is scarce, the verb synthesis figures seem to support to some extent our complexity calculations. The two independent linguistic measures, number of nominal forms and the verbs synthesis figures, correlate weakly (0.33).

For further evaluation we counted mean lengths of word tokens from the Leipzig corpora and correlated these to the complexity figures and noun form numbers with Spearman rank-order correlation. Mean word length can also be considered as a simple complexity measure: the longer the words in the language are, the more there is possibility for variation. Mean word lengths varied from 4.83 (EN) to 8.24 (FI), mean length for all the 16 languages being 5.70 and standard deviation 0.82. Mean lengths of the words correlated with the TTR and MATTR of the Leipzig material highly, 0.88 and 0.91, respectively ($p < 0.001$). Juola complexity figures correlated with the mean length data of the Leipzig corpora more moderately, 0.61 ($p < 0.05$). Mean length of the words correlated highly with the number of distinct word forms: 0.81 ($p < 0.001$).

In Table 5 we show Juola complexity figures of the EU constitution of languages that have available IR data for mean average precision (MAP) improvement achieved by

using some method for word form variation management, such as stemming or lemmatization. The data in the first MAP increase column has been taken from different publications, and is the same as in Kettunen (2009), except for Czech, Portuguese, Italian and Spanish, which originate from McNamee et al. (2009). The data in the second MAP increase column is from McNamee et al's (2009) best n-gramming results, the result being variably achieved with 4- or 5-gramming.

Morphological normalization with stemming, lemmatization or n-gramming reduces morphological variation, i.e. complexity, in an IR database, and thus the IR engine performs better. McNamee et al. (2009) show that the increase in MAP correlates with two measures, namely word length and the Juola complexity measure. Informally this can be seen from the order of Table 5 - gains in MAP are in the same order as the morphological complexity of the language most of the time with only a few exceptions. Only Czech should be below Swedish and Dutch clearly.

	Juola complexity, EU constitution	Best absolute MAP increase Kettunen (2009)	Absolute MAP increase in McNamee et al. 2009 with n-gramming
EN	1.05	2.90	0.09
IT	1.05	4.29	2.5
ES	1.06	4.50	2.1
FR	1.06	3.80	2.9
PT	1.07	3.63	3.6
CS	1.09	10.20	10.24
NL	1.12	5	4.3
SV	1.13	8.80	8.8
HU	1.14	12.40	17.7
FI	1.16	25	16.7
DE	1.17	15.70	9

Table 5. Juola complexity of the EU constitution and best achieved MAP increases for the language with word form variation management

When MAP increase figures are correlated to TTR, MATTR and number of noun forms, we get a view shown in Table 6 with Spearman rank-order correlation co-efficients.

Correlation between	Kettunen (2009) data	McNamee et al. (2009) data
MAP increase and Juola complexity	0.88 $p < 0.001$	0.87 $p < 0.01$
MAP increase and EU constitution TTR	0.85 $p < 0.01$	0.92 $p < 0.001$

MAP increase and EU constitution MATTR	0.92 p < 0.001	0.96 p < 0.01
MAP increase and MATTR Leipzig	0.76 p < 0.01	0.79 p < 0.01
MAP increase and TTR Leipzig	0.94 p < 0.001	0.92 p < 0.001
MAP increase and number of noun forms	0.82 p < 0.01	0.85 p < 0.001

Table 6. Complexity figures correlated with achieved increase in MAP

All the correlations in Table 6 are high and statistically significant or highly significant. These figures support the earlier findings of complexity calculations and give further independent evidence for the plausibility of the calculations.

4. Discussion

In this paper we have shown, that type-token ratio in its basic form and more elaborate sliding window calculation form can be used for approximation of morphological complexity of languages, as has been notified by Juola (1998, 2008) with a smallish data set with typologically more different languages. We have used two different data sets with 21 and 16 languages. We have compared the results of TTR and MATTR to earlier findings with the Juola method with one data set, and given also other evidence that the different calculations correlate most of the time highly. Based on our findings it seems plausible that both TTR and MATTR give a reliable enough approximation of the morphological complexity of languages, even if the used corpora are of very different nature. Specific complexity orders given by the measures vary a bit, but the overall figure seems to conform to linguistic knowledge. Both ends of the complexity scale, least and most complex languages in the sample, stand out clearly, the ones in the middle seem to drift more in different calculations. Specific relative orderings of the languages can be argued about, but this does not weaken the value of TTR and MATTR. They are able to show at least a coarse morphological complexity measure and order for languages. Mutual superiority between TTR and MATTR can not be distinguished based on our data, both seem to be able to show morphological complexity as well.

Some caution with the interpretation of the results is in order. As our other data set, the Leipzig corpus data, is not proper textual material, but random sentences taken out of their contexts, the results achieved with this data are more hypothetical than results achieved with the EU constitution. Anyhow, it seems that the results achieved with the Leipzig data are quite similar with the results of the EU constitution data. It should also be kept in mind, that the EU constitution material represents a quite narrow and constrained textual genre.

Another possible bias in the analysis is that most of the languages in the data are Indo-European and thus more or less closely related to each other and sharing characteristics of their common ancestor (Shosted, 2006). With a larger variation of typologically different languages the results could be somehow clearer.

A relevant next question is, how these complexity figures could be used. As such they do not tell much and it is clear that the complexity figures given by all the three measures do not reveal much about the morphological details of the compared languages. Thus the gained information needs to be used for a quite general purpose linguistic comparison. We'll give a few suggestions on a general practical level.

Firstly, this kind of information could be utilized in approximation of morphological normalization tool needs for IR word form variation management for languages that do not already have existing IR data about the gains that are achievable by morphological normalization (Kettunen, 2013). As more new languages (e.g. languages of India) are being evaluated in an IR setting, this kind of morphological complexity estimation may be beneficial in developing morphological normalization tools for new languages. If you have a sample of new languages for which you want to develop morphological normalization tools, you can estimate the development need and input vs. achieved gain in information retrieval on a general level by comparing the morphological complexity of the languages. The more complex the language is morphologically, the more there is to be gained in IR results by usage of morphological normalization tools (cf. e.g. the situation of English and Finnish/Hungarian). For morphologically less complex languages there is perhaps no need to put so much effort in development of the morphological tools, and quite simple methods can be enough for the purpose (Kettunen, 2013).

Another proposed use in the area of information retrieval is comparison of word form normalization tools, such as lemmatizers and stemmers. If for example several existing lemmatizers or stemmers for a language or several languages need to be compared, their analysis capabilities could be evaluated with usage of the morphological complexity measure they produce out of a corpus. The more the lemmatizer or stemmer reduces morphological variation in the corpus, the less morphologically complex the output of the analysis should be. Although this measure is not in direct relation to achieved IR effectiveness of the morphological tool, it can be useful in the development phase of the normalization tools.

A second suggested application area is machine translation. Statistical machine translation is known to be harder between a morphologically simple and a complex language. For new language pairs morphological complexity can be roughly estimated with the morphological complexity measure and if the languages are very different in their morphological complexity, the translation between these languages should probably use some morphology-aware components, not just statistical information (Sadeniemi et al., 2008).

A final note about the usability of the different complexity measures is in order. If all the three methods we have evaluated can be used to order the languages with respect to their morphological complexity plausibly, then their usage depends partly on practical issues, for example, on the ease of the calculation process. In this respect TTR and MATTR outperform the Juola method, which is more difficult to calculate than type-

token ratio. Especially MATTR is easy to use as an off-the-shelf software, but also basic TTR can be implemented simply.

References

Baayen, R. H. (1996). The effects of lexical specialization on the growth curve of the vocabulary. *Computational Linguistics* 22, 456–480.

Bane, M. (2008). Quantifying and measuring morphological complexity. In *Proceedings of the 26th West Coast Conference on Formal Linguistics*, 67–76. Retrieved 23 January, 2013 from http://www.lingref.com/cpp/wccfl/26/paper1657.pdf.

Bickel, B. & J. Nichols, 2001. Inflectional Morphology. In Shopen, T. [ed.] Language typology and syntactic description Cambridge: Cambridge University Press.

Bickel, B. & J. Nichols, 2004. The AUTOTYP network of typological databases. http://www.spw.uzh.ch/autotyp/

Bickel, B. & J. Nichols. 2005. Inflectional synthesis of the verb. In: Haspelmath, M., M. Dryer, B. Comrie & D. Gil [eds.] The world atlas of language structures. Oxford: Oxford University Press. Available also at http://wals.info/chapter/22

Covington, M. & McFall, J.D. (2008). The moving-average type-token ratio. Retrieved August 29, 2013 from http://www.ai.uga.edu/caspr/Covington-McFall-MATTR-2008poster.pdf

Covington, M. & McFall, J.D. (2010). Cutting the Gordian knot: the moving-average type-token ratio (MATTR). Journal of Quantitative Linguistics 17, 94–100.

Ehret, K. & Szmrecsanyi, B. (to appear). An information-theoretic approach to assess linguistic complexity. In Raffaela Baechler & Guido Seiler (eds.), *Complexity and Isolation*. Berlin: de Gruyter. Retrieved January 25, 2013 from http://www.benszm.net/omnibuslit/EhretSzmrecsanyi_web.pdf.

Hawkins, John A. 2009. An efficiency theory of complexity and related phenomena. In Geoffrey Sampson, David Gil, and Peter Trudgill (eds.), *Language complexity as an evolving variable*, 252-268. Oxford: Oxford University Press.

Hudson, R. (1994). About 37% of word-tokens are nouns. *Language* 70, 331-339.

Iggesen, O.A. (2011). Number of cases. In M. S. Dryer M. and Haspelmath (eds.) The World Atlas of Language Structures Online. Munich: Max Planck Digital Library, chapter 49A. Retrieved January 23, 2013 from http://wals.info/chapter/49A.

Juola, P. (1998). Measuring linguistic complexity: the morphological tier. *Journal of Quantitative Linguistics*, 5, 206–13.

Juola, P. (2008). Assessing linguistic complexity. In M. Miestamo, K. Sinnemäki and F. Karlsson (eds.) *Language Complexity : Typology, Contact, Change*. Amsterdam: John Benjamins Press.

Kettunen, K. (2005). Sijamuodot haussa – tarvitseeko kaikkea hakutermien morfologista vaihtelua kattaa? Master's thesis, Department of Information Studies, University of Tampere. http://tutkielmat.uta.fi/pdf/gradu00702.pdf

Kettunen, K. (2009). Reductive and generative approaches to management of morphological variation of keywords in monolingual information retrieval – an overview. *Journal of Documentation*, 2, 267−290.

Kettunen, K. (2013). Managing word form variation of text retrieval in practice – why language technology is not the only cure for better IR performance? *Trends in Information Management*, 9, 1-21.

Kettunen, K., Sadeniemi, M., Lindh-Knuutila, T. & Honkela, T. (2006). Analysis of EU languages through text compression. In T. Salakoski et al. (Eds.): FinTAL 2006, LNAI 4139, 99–109. Springer-Verlag: Berlin Heidelberg.

Leipzig Corpora Collection Download Page. http://corpora.uni-leipzig.de/download.html.

McNamee, P., Nicholas, C. & Mayfield, J. (2009). Addressing morphological variation in alphabetic languages. In Proceedings of the 32nd Annual International Conference on Research and Development in Information Retrieval (SIGIR-2009), Boston, MA, 75-82.

Moscoso del Prado, M. (2011). The mirage of morphological complexity. Retrieved 20 April, 2013 from http://mindmodeling.org/cogsci2011/papers/0836/paper0836.pdf

Plank, Frans 1986. Paradigm Size, Morphological Typology, and Universal Economy. Folia Linguistica 20, 29-48.

Quasthoff, U., Richter, M. & Biemann, C. (2006). Corpus portal for search in monolingual corpora. In Proceedings of the fifth international conference on Language Resources and Evaluation, LREC 2006, Genoa, 1799-1802.

Sadeniemi, M., Kettunen, K., Lindh-Knuutila, T. & Honkela, T. (2008). Complexity of European Union languages: a comparative approach. *Journal of Quantitative Linguistics*, 15, 185–211.

Shosted, Ryan K. (2006). 'Correlating Complexity: A Typological Approach', Linguistic Typology 10, 1–40.

Sinnemäki, K. (2011). Language universals and linguistic complexity. Three case studies in core argument marking. General Linguistics, Department of Modern Languages, University of Helsinki. Retrieved 20 April, 2013 from https://helda.helsinki.fi/handle/10138/27782

Stump, G. T. (2001). Inflection. In A. Spencer and A. Zwicky (eds.), *The Handbook of Morphology*, 13–43. John Wiley and Sons: Hoboken, NJ.

Tweedie, F. J. & Baayen, R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, 32, 323–352.

Acknowledgements

Most of this paper was finished while the author was visiting UFAM, Universidad Federal do Amazonas, Institute of Computing, and funded by FAPEAM, Fundação de Amparo à Pesquisa do Estado do Amazonas (http://www.fapeam.am.gov.br/) with grant number 159/2012.

The author wishes to thank anonymous referee of J. of Quantitative Linguistics for useful comments when preparing the final version of the paper. Prof. Johanna Nichols kindly provided the AUTOTYP verbal inflection synthesis database for use.