

## Summary

The objective of the study was to demonstrate, whether a relation existed between the lexical diversity of patent documents. In principle the supposition was that the richer the lexical diversity the higher the possibilities of the patent being granted. There were a few clues for that relation could to be logical, first that a patent agent having a deeper knowledge of the technology would probably use a richer vocabulary and the other that a document written with a better English could reflect not only a better command of the language, but the logical rules that the language allows to use.

Four lexical diversity indexes were obtained from the abstracts of 450 patent documents filed by applicant of nine non English speaking European countries.

The initial comparisons between the granted and non granted patents didn't show any difference in lexical diversity. However, when the patents under the 20% lowest percentile in the indexes results were compared with the patents above the 20% highest percentile a surprising result (unless you are a patent examiner, I suspect) appeared.

A Chi square test was performed to assess if there is an association between the variables Lexical Index (low, high) and Granted patent (0 = No, 1 = Yes). In the four cases a confidence level of 95% didn't allow to reject the lack of association hypothesis, however the p-values obtained from the calculations showed that the Type-Token Ratio Index (TTR) and Herand C indexes lack of association hypotheses would be rejected, if a confidence level of 90% was adopted.

The t-tests comparing granted and non granted patent means gave up a similar result. In this case the TTR t-test comparison gave out a p-value of 0.05667, very close to the limit of 0.05 below which the granted and non granted patents would show different means in lexical diversity. The Herand C index when the two subsets of the data were subjected to the t.test gave a p-value of 0.08 what is, as with the Chi square results not enough to reject the hypothesis that the means are identical under a 95% confidence level, however both hypotheses would be rejected if a 90% level was used.

The surprise came when comparing the four lexical diversity means of the granted patents with the means of the non granted. Granted patents presented in the four cases lower means than non granted patents. That coincidence and the fact that with a 90% confidence level the difference of means and the association of variables hypothesis would be accepted in two of four indexes suggests that the effect of using very complex vocabulary in the patent's text is negative on the chances of being granted.

There might be a quite logical explanation for that result. Since patent examiners have to read many applications it is probable that they feel more inclined to give a

positive opinion when reading an application written with a very simple language than when reading one with a very complex vocabulary. Imposing patent examiners an unusual effort to read a patent application might then not be a good idea however, it is highly unlikely going to be the main reason why the patent application wasn't issued/granted.

### **The data**

The dataset consist of 9 sets of data that were downloaded from the EPO database, accessed via the patent search engine Global Patent Index (EPO Patent information services for experts) Each original set of data contained 50 patents filed by European applicants from a non English speaking country to the USPTO The patents had been published between 2005 and 2008 in the United States of America, taking into account that the average patent processing time in USPTO is of four years by the day in which the dataset was obtained the vast majority of them will have reached an issued (granted) or non issued decision. The search string for Germany was: ((PRC AND APPC) = DE) AND (PUC = US) AND PUD>20050101 AND PUD<20080101 Where PRC means priority country, APPC means applicant country, PUC publication country and PUD publication date

The dataset and the R codes used do transform and analyze the data are stored in the github platform and can be freely accessed via the link:

<https://github.com/WernerDJ/Patents-and-Lexical-diversity>

## The lexical diversity distributions

Index	Minimum	Median	Mean	Maximum	SD
TTR	0.3099	0.6290	0.6430	1.0000	0.1403 2
Herand C	0.7251	0.8890	0.8899	1.0000	0.0474 7
Guiraud R	2.611	4.953	5.003	8.393	0.9590 9
Uber Index	6.735	16.490	20.502	180.905	15.256 1

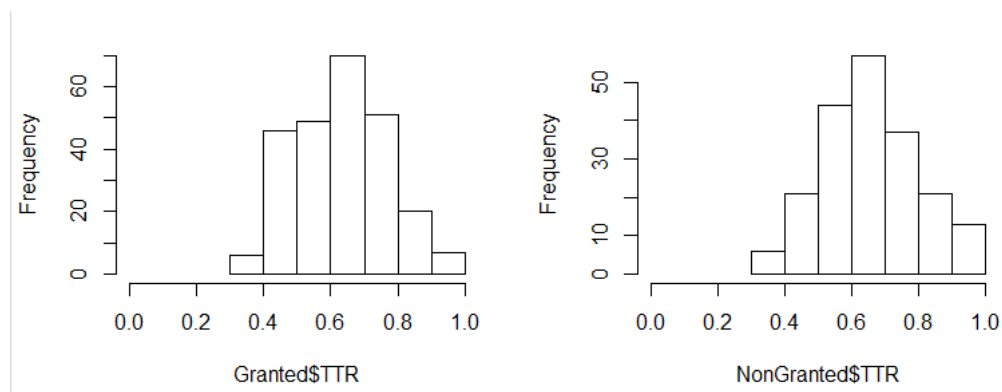
### TTR

Shapiro Normality test

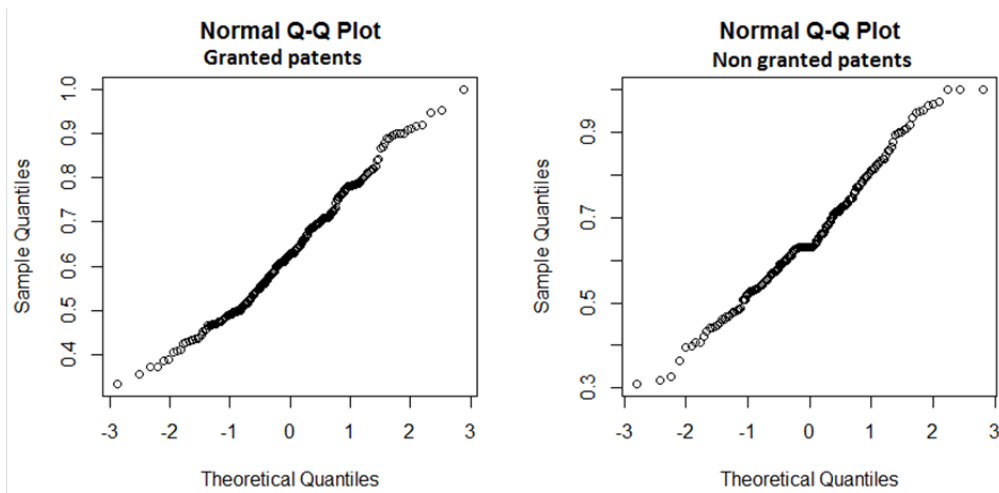
Granted patents p-value: 0.02253

Non Granted patents p-value: 0.05392

The Granted patents do not follow a normal distribution however, the non granted seem to do so, although the p-value is very close to the limit 0.05



The difference in the normality test result is quite strange, especially taking into account that the histogram plots, in both cases look like normal distributions. Another test the Q-Q normal plot, visual as well, plots the distribution of data values against a vertical axis in which the distribution of normal quantiles is represented. If the data follows a normal distribution the points follow a straight diagonal line across the graphic.



In either Granted and non granted patents the TTR distribution follows very closely a normal distribution

#### Herdan C

Shapiro Normality test

Granted patents p-value: 0.1579

Non Granted patents p-value: 0.07579

In both cases (granted and non granted patents) normality cannot be rejected

#### GuiraudR

Shapiro Normality test

Granted patents p-value: 0.4138

Non Granted patents p-value: 0.5529

In both cases (granted and non granted patents) normality cannot be rejected

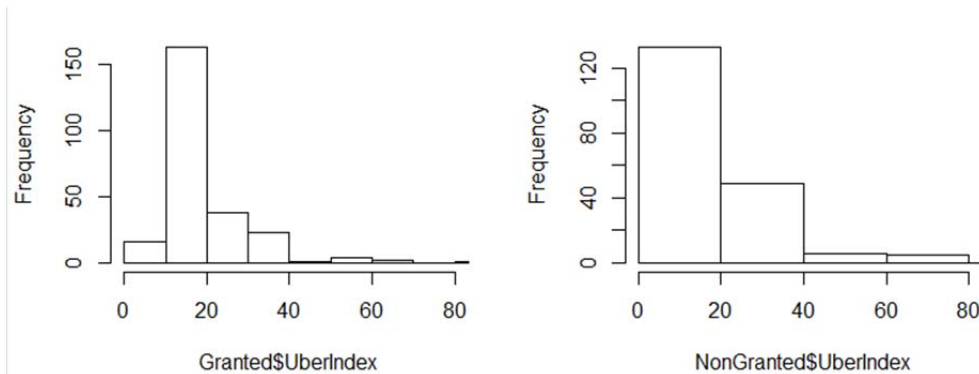
#### Dugast's Uber Index

Shapiro Normality test

Granted patents p-value: 2.2e-16

Non Granted patents p-value: 2.2e-16

The distribution of both granted and non granted patents do not follow normal distributions



The histogram graphic of the Dugast's Uber Index (left graphic) show what look like exponential distributions a logarithmic transformation of the frequencies show a linear descending trend (right graphic) as is usual when a power distribution is transformed by logarithms.

### T test (Welch Two Sample t-test) with a 95% significance level

H0 – The means of granted and non granted patents are equal

HA – The means are different

	t-values	Degrees of freedom	p-values
TTR distribution	1.6442	410.45	0.1009
Herdan C	1.6313	414.92	0.1036
Guiraud Root	0.75596	418.64	0.4501
Uber Index Wilcox test (non-parametric)	W = 25926		0.227

The next step was to further develop the initial hypothesis of an influence of the vocabulary diversity by comparing the patents with the lowest vocabulary diversity with those with the highest. The cutoff was arbitrarily set in the 20% lowest and highest indexes. Four datasets, one for each variable were constructed.

Since the new data had been divided in four categories of two variables, low vs. high (variable of Lexical index value) and granted vs. non granted (outcome variable of the patenting process) the Chi square test for 2x2 tables was the natural first step to analyze if there was an association between both variables:

#H0 No association between the variables Lexical Index and patenting outcome

#HA there is association between both variables

Herdan C

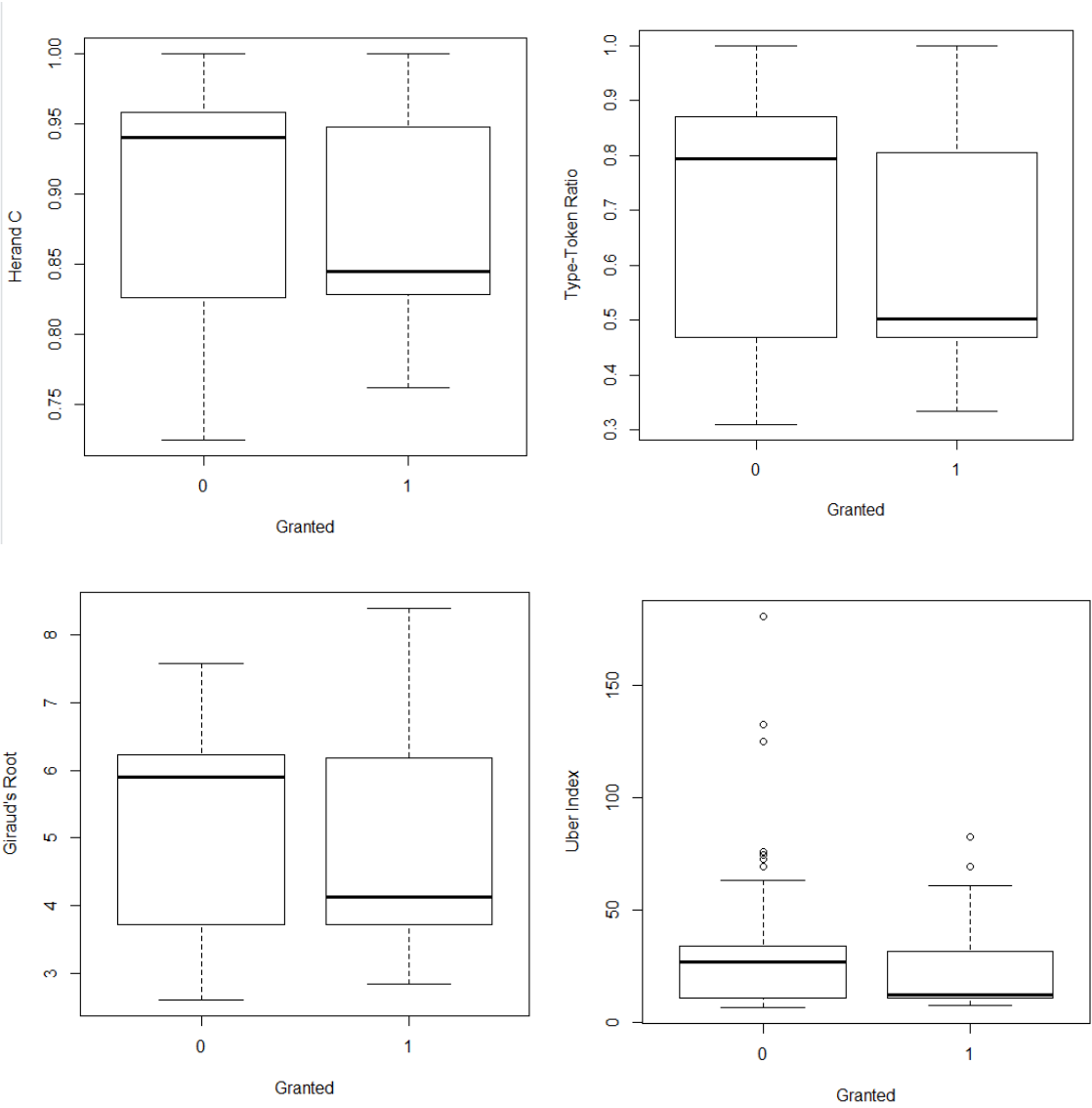
	Low	High
Non Granted	44	32
Granted	44	56

A second test was run on the four to see if there was a difference in the vocabulary index means or distributions of the granted and non granted patents

Dataset	test	p-value	Refute H0 with 95% confidence	Refute H0 with 90% confidence
Herdan C	Chi square	0.0678 <sub>3</sub>	NO	YES
	t-test	0.0838 <sub>2</sub>	NO	YES
Type-Token Ratio	Chi square	0.0658	NO	YES
	t-test	0.0566 <sub>7</sub>	NO	YES
Giraud's Root	Chi square	0.8792	NO	NO
	t-test	0.7856	NO	NO
Dugast's Uber	Chi square	0.1292	NO	NO
	wilcox.test	0.2622	NO	NO

Finally a summary of the means in granted and non granted patents of the four used indexes showed that in all cases the lexical diversity mean of granted patents was lower than in non granted patents.

Index ordered data	Mean of the non granted patents	Mean of the granted patents
Herdan C Index	0.8995965	0.8806790
Type-Token Ratio	0.6899532	0.6278951
Giraud's Root Index	5.049695	4.990261
Dugast's Uber Index	30.09606	22.32343



## **Conclusion**

It has to be taken into account that the data analyzed was the 20% lower and higher index results. The t-tests and the means differences suggest that a highly complex vocabulary has an effect on the chances of a patent being granted, however the effect is slight. If a patent in which a complex vocabulary was used is not granted most of the blame will be due to the causes pointed out by the examiner and at the same time writing a very easy to read patent will not constitute a guarantee for the issuing / granting of the patent application. There is however one clear advice that should be taken into account in view of this study: Keep it simple. Writing a patent application that is as easy to read as the technical matter makes possible might help not only the patent examiner but even yourself.