BILETA
British & Irish Law, Education and Technology Association
2007 Annual Conference
Hertfordshire 16-17 April
UH

A quantitative analysis of legal word use

Kees van Noortwijk and Richard De Mulder
Centre for Computers and Law
Erasmus University Rotterdam
Email: vannoortwijk@frg.eur.nl & demulder@frg.eur.nl

Abstract

Jurimetrics, the empirical study of legal phenomena, is generally associated with the quantitative analysis of legal cases. However, an interesting field, which can also be seen as a part of jurimetrics, is the analysis of legal word use. The subject matter of this examination was a broad selection of British legal documents, both case reports and legislation. Using quantitative methods developed by linguists, these documents were compared with each other but also with documents containing ordinary English usage. These three different types of documents were roughly the same size: about 16 million words each. This paper presents the preliminary results of this examination. It also presents the results of an exploratory comparison between the word use in English legal documents/ordinary English documents and that between Dutch legal documents/ordinary Dutch documents. These results could help to improve the accuracy of legal retrieval systems.

Introduction

In the formulation, interpretation and formulation of law, language almost always plays an essential role. It is the medium, the process and the product in all kinds of legal activities.[1] An important task in legal knowledge management, and indeed in almost any type of legal work, is therefore the management of legal documents. However, the retrieval of documents stored in electronic repositories (including commercial databanks) has proved to be far more complicated than expected, and the level of accuracy in the retrieval of relevant documents is not always satisfactory.

This paper argues that legal document retrieval can be improved when more is known about the characteristics of the documents. Some of these characteristics, such as added 'metadata' describing the subject matter of documents, are already widely used by retrieval systems. One characteristic that is not commonly used, however, is the *word use* in these documents. It is commonly presumed that the word use in legal documents is different from that in other documents. If this is indeed correct, that is if certain types of legal documents – the texts of statutes and case reports – would contain different words or would show different word use patterns than 'general' texts, such characteristics might be useful to improve retrieval systems.

A pilot study on this subject, carried out with respect to the Dutch language a decade ago[2], showed some interesting results. For instance, word frequency distributions (the 'pattern' of word use) proved to be quite different in legal texts than in 'general Dutch'. It was, therefore, interesting to investigate whether the same would hold true for document comparison in other languages and whether there would be differences between Dutch legal texts and texts in other languages. It was decided to examine British English texts. Two corpora containing thousands of legal documents (one containing legislation and one containing case reports) and a corpus containing general texts, all in British English, have been compiled. The corpora are of roughly equal sizes (around 16 million words each). The 'general' corpus consists of a random sample from the 'British National Corpus'[3], the two legal corpora consist of legal texts available on the internet. Cases for the case reports corpus have been

[1] Maley (1994, p. 11).
[2] See Van Noortwijk (1995).
[3] The British National Corpus (BNC) is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of current British English. For this project, only the written sources were used. See http://www.natcorp.ox.ac.uk/.

selected in such a way that the percentage of cases heard by various courts in the British hierarchy of courts (House of Lords, High Court, County Court cases etc) more or less corresponds to that in the ten year old Dutch case law corpus, which will facilitate inter-language comparison.

Using these corpora, research was carried out to map as precisely as possible the differences between word use in the respective language types. The results were then compared to those from the Dutch language study. Several of these preliminary results are presented in this paper. A report containing the complete results will appear later this year. It is expected that results from this research project will be of importance for the development of new, more 'intelligent' legal document retrieval systems in the near future.

Jurimetrics: empirical legal research

It was the American Lee Loevinger who launched the term "jurimetrics"[4]. He stressed the importance of scientific, and therefore quantitative methods for lawyers. The main stream in jurimetrical research has always been based upon the work of the North American "legal realists". Yet apart from the American publication "Jurimetrics Journal", the work of Stuart Nagel and the occasional contribution by Ejan Mackaay, the jurimetrics front has become rather quiet since the last legal realist (probably Reed Lawlor[5]) stopped his research activities. A few years ago, a trend developed to use jurimetrics research in order to produce legal advice systems. Smith and Deedman, and John Zeleznikov and Dan Hunter are the names that come to mind in this connection. Perhaps it is fair to contend that it is the Netherlands where jurimetrics has enjoyed the most interest. Academics such as Franken, de Wildt, De Mulder, Van Noortwijk, Piepers, Combrink-Kuiters, Snijders and Malsch have devoted their academic careers to this subject.

Before proceeding any further, it is useful to define what is meant by the term Jurimetrics.

> *Jurimetrics is concerned with the empirical study of the law in the widest sense, not only the meaning (the semantics) but also the form and the pragmatic aspects of law. Law is defined here as the demands and authorisations issuing from state organisations.*

It is the empirical, quantitative and economic approach to law that will enable lawyers to give advice that will be relevant, reliable and comprehensible to their clients. For instance, lawyers can then provide their clients with reliable and valid estimates of legal risks and costs. These factors can often be calculated if the enormous amount of legal data, in particular the case law that is now available in electronic form, is studied in a jurimetrical way. This involves the analysis of various aspects of legal language. A common characteristic of many publications on this subject is that they concentrate on *semantic* aspects, i.e. aspects which have to do with the *meaning* of the text. An example is research which concentrates on ascertaining which are the most important objectives of a certain statute. This is often connected to questions of a pragmatic kind, for example whether these objectives could be achieved in practice. In addition, attention is sometimes paid to the *form* legal texts have, but in most cases this is studied in relation to the meaning. For example: because of long or unusual words or because of a difficult sentence structure, legal texts are difficult to understand.

However, studying the form of legal texts can be interesting in itself without dealing with the semantic aspect of the text as well. The pilot project on quantitative aspects of Dutch legal language described in [Van Noortwijk 1995] already yielded some interesting results in this respect. It was found, for instance, that the word use in legislation differs measurably from that in the texts of case reports, and that both differ significantly from general Dutch texts. Having obtained detailed knowledge on this subject, an interesting question is, of course, if this knowledge can be put to use for the legal field. This question is addressed briefly below.

Quantitative linguistics

When a human reader, even one who does not have any legal knowledge, compares the texts of statutes or case reports to other texts, it is quite likely that he will notice differences in form. For example, more formal words are used, sentences are structured in a different way, and certain headers

[4] Loevinger (1949).
[5] Lawlor (1967), Ulmer (1967) and Goldman (1971).

are used. In many cases these differences are so obvious that it is possible to distinguish the legal from a general text type just by glancing at them for a few seconds. This being the case, an important question from the scientific point of view is if these differences in form can in some way be made concrete and expressed in figures. To address this question we have used several methods from the field of *quantitative linguistics* (sometimes also referred to as *statistical linguistics*)[6]. This is a branch of linguistic science in which the measurability of linguistic phenomena plays a central role. We will limit ourselves to the *word use* in texts for the moment and ignore characteristics such as the structure and the length of sentences or longer text entities.

Carrying out research into the word use in certain text types used to be very labour intensive. To acquire a good overview, it is often necessary to study large text files or *corpora*. Word lists of these corpora had to be compiled manually until the early 1960's. Since then, computers have gradually taken over this task. Nowadays, practically all recent text material is available in electronic form because of 'electronic publishing' and electronic typecasting. Fast computers are available to many researchers. Word counting and syntactic analysis of text material can now be done in a relatively short time. This makes it possible to compile and compare very large text corpora, for instance a corpus containing every statute that is in force in a certain country at a certain moment. Some limitations still exist. Even at the syntactic or form level it is sometimes desirable to be able to understand the *meaning* of a text. However, in most cases computers are not capable of this and are unable, therefore, to render the meaning of the words they count. Consequently, some characteristics of a word are difficult to record, such as the lexical category to which it belongs. We have avoided this limitation by using nothing but the original word forms from the text corpora. This means that we have not tried to reduce a word to its 'stem' form (lemmatisation) or to distinguish different lexical categories.

Quantitative linguistics provides a number of methods to analyse the word use in a corpus. Characteristics which play an important role in these methods are for instance *word frequency* (how often does a certain word appear), *frequency distribution* (what is the pattern of the word frequencies of all the different words in a corpus) and *distribution of word types* (whether a certain word used in every document in a corpus, or only in a subset of documents). These characteristics can be analysed by compiling a *frequency list* of the corpus. This is a list of all the different words (or 'word types') in the corpus, plus the number of times the word appears in the corpus and the number of documents of which it is a part[7]. This list is sorted according to word frequency, the most common word being at the top. Word types are give a *rank number*, based on their position in this frequency list: 1 for the most common word, 2 for the next most common word, etc. Based on this frequency list, a number of linguistic measurements can be made, such as the 'characteristic *K* of Yule/Herdan (explained below)[8]. The value of these measurements provides a typology of what could be called the 'structure of word use'. Apart from these data, which characterise the *way in which words are used*, the words themselves have also been studied in this research project. Points that have been taken into account in this respect are, for instance, *word lengths* and the specific words which appear at the 'head' of the frequency list (the most common words in a corpus).

Legal corpora

The characteristics mentioned above can be used to record certain quantitative data or measurements about a text corpus. To apply these characteristics, however, suitable corpora must first be available. These corpora must then be given an appropriate form. Although legal language has both a written as well as a spoken form, for the purposes of this analysis two specific written forms were selected, namely legislation and case law reports. Both form important sources of law, each with its own characteristics.

Of interest here is whether the language in these two types of documents differ measurably: is the language used to produce statutes different from the language found in case reports? Furthermore, is it different from the language in other (non-legal) documents? To answer these questions, two separate legal corpora – sets of documents forming a single data collection – have been compiled, together with a corpus containing general English. To facilitate comparisons between these corpora, it

[6] See for instance Guiraud (1959) and Herdan (1966). For an overview of the developments in the field of quantitative linguistics, see Bailey (1969) and Baayen (1989).
[7] See for examples of the use of these characteristics for instance Kucera & Francis (1967).
[8] Van Noortwijk (1995, p. 27).

was ascertained that they were of roughly the same sizes and were constructed according to the same standards (i.e. no differences in the form of special codes, characters etc). Compiling these corpora was not a small task. This section contains a description of the different steps involved in this process and the complications that were encountered.

Legislation

In the early 1990's, when the material was gathered for the Dutch language pilot project, it was rather difficult to obtain a sufficient number of statute texts and regulations. Eventually this was solved by the cooperation of the publisher of the major Dutch legislation databank. Combining the databank records into documents of useable size was quite a task, however. Currently, most legislation, whether from The Netherlands or the United Kingdom, is published on the Internet. The BAILII[9] web site served as the main source of these legal texts for this project.

In the summer of 2006, an inventory was made of the primary and delegated legislation available on this web site. It was decided to take a random sample from these texts. The sample had to be as large as possible, but it had to contain roughly the same percentage of primary legislation (acts of parliament) and delegated legislation (in this case, 'statutory instruments') as was the case in the 1995 Dutch pilot project, namely 20% vs. 80%. This restriction was made in order to optimise the comparability between results from the pilot and results from this new project. After careful consideration of the list of available documents, taking into account the size in words of each document in order to apply the restriction mentioned above, a total of 3109 documents (627 acts of parliament and 2482 statutory instruments, roughly a 1:4 ratio), from the period 1972-2006, were selected and downloaded.

The documents from BAILII were in HTML format, i.e. they contained HTML mark-up. The first task was to remove this mark-up so that only the 'pure text' of the documents remained. This was achieved by means of a program called 'Web2Text', which does exactly what its name suggests. When the resulting text files were inspected, however, some remaining errors were discovered. Several BAILII-specific separation codes, headers and footers were still present and had to be removed. Furthermore, many words proved to be broken down into strings of characters or even single characters by the insertion of white space or line breaks. To fix this, a series of automatic and manual corrections had to be made. This finally resulted in a corpus containing just over 16.5 million word tokens.[10]

Preparing this corpus for further analysis involved two more steps. First, every corpus document was converted into a 'word type file', a file listing all the different words in the original document together with their frequency (the number of times they appeared in the original document). Table 1 contains a fragment from one of these word type files (sorted by word frequency):

any,23
and,21
expenses,20
person,20
regulation,19
distributor,19
...

Table 1 - Fragment from one word type file

As a second step, the separate word type files were combined into one new file, containing all the word types from the corpus together with their total word frequency (the frequencies from the separate files added) and with their 'document frequency', the number of documents they had appeared in. Table 2 shows a fragment of this list:

[9] BAILII, the British and Irish Legal Information Institute, http://www.bailii.org.

[10] A 'word token' is a string of characters, separated from other strings of characters by one or more 'separation characters' (white space, punctuation, etc.). A word token can appear more than once in a single document. As opposed to this, the term 'word type' is used to indicate the *different* words that appear in a document (equivalent to the 'vocabulary' present in that document). A corpus consisting of 1000 word tokens could, for instance, contain 100 (different) word types.

paragraph,2130,79240
not,2035,78039
such,1709,73167
order,2302,63694
it,2092,63579
3,2662,62752
...

Table 2 - Fragment from the combined word type file

In this table, the first number after the word type is the number of documents in which the type appeared and the second number is the total word frequency. Numbers, such as '3' in table 2, are treated like any other sequence of characters and are therefore stored in the type list just like any other character or word.

To compile these lists, two software applications were developed, one to make the word type files and one to combine them. These applications also store relevant statistical information documents and word types that are processed.

Case law

The case law corpus was compiled in a similar way. In the Dutch pilot project, a commercial case law database[11] had been used, which contained a mixture of civil and criminal law cases from various courts[12]. In this new project, an attempt was made to build a corpus that was of more or less the same size as the legislation corpus (16.5 million tokens) and contained the same percentages of civil law and criminal law cases (and a comparable mixture of court types) as in the pilot project. This resulted in the selection of a set of cases as specified in Table 3.

Court	Number of cases
England & Wales Court of Appeal, Civil Division Decisions	1381
England & Wales Court of Appeal, Criminal Division Decisions	540
European Court of Human Rights Decisions	17
United Kingdom House of Lords Decisions	108
Scottish High Court of Justiciary Decisons	423
Scottish Court of Session Decisions	153
Scottish Sheriff Court Decisions	85
England and Wales High Court (Admiralty Division) Decisions	3
England and Wales High Court (Chancery Division) Decisions	126
England and Wales High Court (Commercial Court) Decisions	96
England and Wales High Court (Sup. Court Cost Office) Decisions	9
England and Wales High Court (Family Division) Decisions	20
England and Wales High Court (Patents Court) Decisions	14
England and Wales High Court (Queen's Bench Division) Decisions	57
England and Wales High Court (Techn. and Construction Court) Decisions	40
Total	3072

Table 3 - Composition of the case law corpus

The House of Lords and Court of Appeal cases roughly account for 52% of the corpus, the High Court (all divisions) together with the Scottish Court of Sessions and Scottish High Court account for 44%, the Scottish Sheriff Court for 3% and the European Court of Human Rights for 0.5%. Although this might not be a perfect reflection of the actual number of decisions that is issued in a certain period in each division, the set comprises a broad selection of cases while its composition should make it possible to compare quantitative results with those from the Dutch pilot project.

[11] This was the electronic archive of the leading Dutch case law report 'Nederlandse Jurisprudentie', containing 16430 cases from the period 1965-1989.

[12] At the time, the contents of the Dutch databank 'Nedelandse Jurisprudentie' that was used for the pilot project contained 52% High Court decisions (Civil and Criminal law), 21% Court of Appeal decisions, 23% District Court decisions, 3% First Instance Court decisions and around 0.5% European Court of Human Rights decisions.

After this selection procedure, the appropriate documents were downloaded from BAILII. Many of the case reports were found to be available as RTF text files this time. If that was the case, we used this RTF version, as this is much easier to convert to a 'pure text' file. If not, we used the HTML-version and converted it using Web2Text. After inspection – and where necessary correction – the documents were converted to type files again and a type file for the whole corpus was created.

General language: British National Corpus

In the Dutch pilot projects, it had been rather difficult to find a suitable 'general language' corpus. This was eventually solved by using a custom-made corpus from the Institute of Dutch Lexicology. For the English language, this was easier. The British National Corpus, maintained by the University of Oxford, is well suited for the type of comparisons required. It is "a general corpus that includes many different styles and varieties, and is not limited to any particular subject field, genre or register. In particular, it contains examples of both spoken and written language."[13] A licensed version of the corpus was obtained, from which, in this case, only the part representing written language was used.

The full corpus (both written and spoken language) is over 100 millions words in size. For this project, it was decided to take a 16.5 million word (approximately) sample of written language from the corpus. The reference file that comes with the corpus was used to make a selection. This file contains details about every document in the corpus (the documents themselves are stored as individual files), such as the size and the type (or 'genre') of the document. The corpus manual contains a list of all different text types. It was decided to leave out documents with the following types:

- spoken text;
- written to be spoken;
- texts before 1975;
- texts made for children;
- legal texts.

The reason for leaving out older documents (those from before 1975) was that the case law corpus contained only case reports from 1975 onward. It was therefore decided to limit the legislation corpus to statutes passed after 1975 in order to avoid archaic language and uncommon words (that is, uncommon in recent documents). Leaving out the texts made for children was necessary to develop a more homogenous corpus whereas removing the legal texts was done because the corpus was to be used as the counterweight to the legal texts, to find *differences* between general and legal language, which might be obscured if legal language would be present in both.

After the removal of these types, some 82.5 million words remained. A random sample of roughly 20% of this remaining part was taken and the corresponding documents were copied to a separate folder. The XML mark-up in the files was removed again using Web2Text. Then a header text, included by the editors of the BNC, was also removed after which a word type file was created for every separate document, together with a word type file for the whole corpus.

First comparisons

Some statistical data about the three corpora, gathered while compiling them, are given in table 4. As these data show, the corpora differ considerably, even when we limit ourselves to the most basic characteristics. The total number of word tokens is of course practically identical, this was the size criterion used to decide how many (complete) documents to include in each corpus. The number of word types – the vocabulary – differs considerably, however. The overall token/type ratio varies from 173.5 (legislation), via 134.9 (case law) to 72.6 (BNC). This means that on average, every word type is used more than twice as often in both legal corpora than in other text types.

As for the size of the documents, it can be seen that most documents in the BNC corpus are considerably larger than in both legal corpora. *Differences* in sizes (expressed by means of the standard deviation) are also the highest in the BNC corpus and by far the lowest in the case law corpus.

[13] See http://www.natcorp.ox.ac.uk/corpus/.

Corpus	UK Legislation	UK Case law	BNC Corpus
Size in bytes	104414786	98155538	102144099
Total number of tokens	16549886	16500062	16885023
Total number of types	95384	122354	232684
Number of documents	3109	3072	611
Largest document			
Size in bytes	2925923	420685	1148389
Number of tokens	489222	71206	192283
Number of types	8859	5859	15193
Smallest document			
Size in bytes	116	55	1860
Number of tokens	15	5	288
Number of types	14	4	170
Average			
Size in bytes	33584.69	31951.67	167175.28
Number of tokens	5323.22	5371.11	27635.06
Number of types	468.77	964.36	4222.51
Standard deviation			
Size in bytes	115089.23	39371.26	128937.50
Number of tokens	18827.62	6656.64	21093.88
Number of types	720.63	664.94	2455.38

Table 4 - Desciptive statistics

Linguistic constants

Researchers in the field of quantitative linguistics have proposed several characteristics (measurements) to compare corpora. One of these is the so-called 'Characteristic *K*', as defined by Yule and Herdan. This is an indication of the average frequency of the repetition of word types. According to Yule, at least, this characteristic is therefore also an indication of the size of the vocabulary in a corpus (i.e. it could be used to predict the number of word types from any given number of word tokens). In the pilot project it was found that it appears to be sufficiently stable in samples of different size, taken from a corpus.[14] *K* can be calculated as follows:

$$K = \frac{\sum r.n_r}{\left(\sum r^2.n_r\right)^2} \qquad (1)$$

where *r* stands for the rank number of a frequency class, equal to the frequency in the corpus of the word types in that class, and n_r stands for the number of word types in the class. These values for *K* are calculated for the three corpora:

Corpus	*K*	*K* (Dutch pilot)
Legislation	0.0134	0.0128
Case law	0.0135	0.0111
BNC	0.0090	0.0106

Table 5 - Characteristic *K* for the different corpora

The two legal corpora yield values that are almost identical, whereas the general language corpus yields a *K* that is considerably lower. In the Dutch pilot project similar results were found, but with an in-between score for the case law corpus. An explanation could be that the word use in British case law reports is rather formal – much like that in legislation texts –whereas Dutch case reports might

[14] See Van Noortwijk (1995, p. 87).

contain a mixture of formal (legal) language and more general discourse. More research will be needed to see if that is indeed the case, however.

Another relationship between the number of word tokens and word types, as defined by Erikstad[15], has also proven to be relatively stable, no matter what the size of a corpus. In this relationship, the number of word types in a sample is considered equal to a power *C* of the number of word tokens in the sample, multiplied by a constant *R*.

$$V = R.N^{c} \qquad (2)$$

By using known values for the numbers of tokens and types in different samples, the values of *C* and *R* can be calculated by means of regression analysis. The results, for the British corpora as well as for the Dutch pilot, are given in Table 6.

Corpus	**British corpora**		**Dutch corpora**	
	R	***N***	***R***	***N***
Legislation	10.79	0.55	13.30	0.57
Case law	7.24	0.59	9.83	0.58
BNC	21.81	0.56	6.91	0.65

Table 6 - Token/type ratio constants

Explaining the different values is beyond the scope of this paper. Interestingly, the British National Corpus shows different values for *R* and *N* than perhaps could be expected from the pilot project. Because the square of the Pearson product moment correlation for the regression analysis underlying the values of these constants is very high (above 0,998 for all corpora), the reliability of formula (2) for the estimation of the number of word types in a sample of any given size is also high.

Word frequencies

When studying the frequencies of the word types present is the corpora, an interesting feature is the relationship between the rank number of word types and their frequency. For instance, figure 1 shows a chart in which the *cumulative frequencies* of the word types in the corpora are plotted against their respective rank number. Especially when the rank number is plotted using a *logarithmic* scale, differences between the corpora are clearly visible. In general, the pilot project already showed that logarithmic ratios often play an important role with language-related phenomena. An explanation for the differences that are visible might be that in the legal corpora, the high frequency word types (specifically the ones in the top 10 of most frequent words) account for a higher number of tokens than is the case in the general text corpus. In other words, these very frequent words are used even more frequently in legal documents than in other documents.

[15] See Erikstad (1980, p. 223).

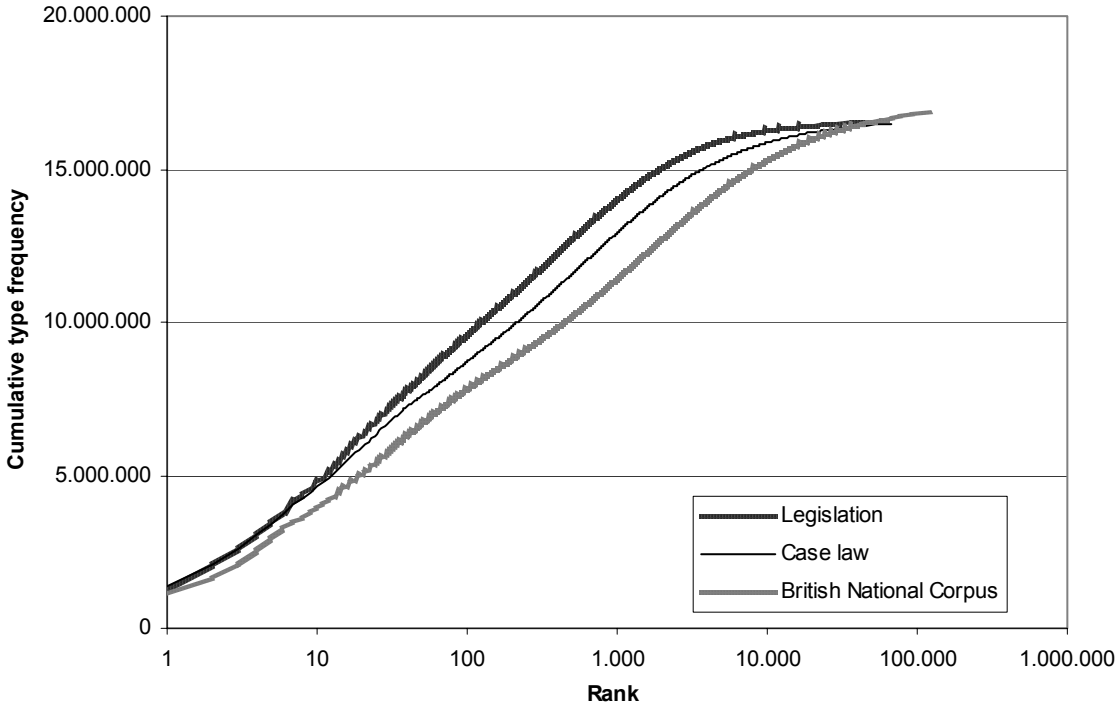

Figure 1 - Cumulative frequencies by rank number, log. scale

Frequency distributions

A frequency distribution is a list of all word frequencies that are present in a corpus, with each frequency being coupled with the number of word types that have frequency in the corpus. This list is usually sorted by frequency, the most common frequency (most of the times equal to 1) at the top. When the frequency data are laid out in this way, an impression can be obtained of the *structure* of the word use. However, for large corpora a problem is that the frequency distribution can be very bulky and therefore difficult to survey. In the British corpora, for instance, more than 2,000 different frequencies are present. Therefore, it is better to use *classified* frequency distributions. These are now being created for these corpora and will be the subject of future publications on this subject.

Conclusions

The aim of the project described in this paper is to conduct a thorough analysis of all measurable characteristics of English legal word use and to compare this with English general word use. Once this analysis is complete, it will be possible to compare the results with those that have already been obtained for the Dutch language. Although it could be presumed that there would be differences between English legal texts and ordinary legal texts, what was of interest here was to know what these differences were and whether they corresponded to those that had been found in the Dutch language project.

What emerged from the Dutch language project was that the structure of word use in legal corpora differ at certain points from that in the general Dutch corpus. It was found that in legal corpora fewer word types are very common, whereas each of these word types has on average a higher frequency as well. There are some differences between the legal corpora as far as the structure of word use is concerned as well. The differences between the corpora can be verified by means of certain linguistic constants. Of these, the characteristic K of Yule and Herdan and the ratio $V = R.N^c$ of Erikstad in particular seem to be more or less insensitive to the size of the corpora.

Studying the word use in legal documents provides empirical knowledge about the contents and the structure of these documents. A number of techniques, many of which have their origins in quantitative linguistics, can be used in this work. Results from this research could eventually lead to the development

of more powerful legal information systems. The statistical analysis as presented in this article belongs to the field of jurimetrics, as it is an empirical and quantitative analysis of legal phenomena. It is different from most jurimetrical studies in the sense that it looks at the form of legal texts rather than their meaning and pragmatics. It is an area that has perhaps been overlooked until now.

References

Baayen, R.H. (1989), *A corpus-based approach to morphological productivity*, Dissertation, Amsterdam: Centre for Mathematics and Computer Science (Free University) 1989.

Bailey, R.W. (1969), 'Statistics and Style: a historical survey', in: Dolezel, L. & Bailey, R.W. (eds.), *Statistics and Style*, New York: Elsevier 1969, p. 217-236.

Erikstad, O.M. (1980), 'Appropriate Document Units for Text Retrieval Systems', in: Bing, J. & Selmer, K.S. (Eds.), *A Decade of Computers and Law*, Oslo: Universitetsforlaget 1980, p. 220-238.

Goldman, S. (1971), 'Behavioural approaches to judicial decision-making: Towards a theory of judicial voting behaviour', in: *Jurimetrics Journal*, March 1971, p. 142.

Guiraud, P. (1959), *Problèmes et Méthodes de la Statistique Linguistique*, Dordrecht: Reidel publishing company 1959.

Herdan, G. (1966), *The advanced theory of language as choice and chance*, Berlin: Springer-Verlag 1966.

Jensen, M.C. & Meckling, W.H. (1994), 'The Nature of Man', in: *Journal of Applied Corporate Finance*, Summer 1994, V. 7, No.2, pp. 4-19.

Kucera, H. & Francis, N.W. (1967), *Computational analysis of present-day American English*, Providence: Brown University Press 1967.

Kuhn, T.S.(1970), *The Structure of scientific revolutions*, Chicago/London 1970.

Lawlor, R. (1967), 'Personal stare decisis', in: *Cal. Law Rev.*, vol 73, p.41.

Loevinger, L. (1949), 'Jurimetrics, the next step forward', in: *Minn. Law Rev.*, april 1949, p. 455.

Maley, Yon (1994), 'The language of the Law', in: Gibbons, J., *Language and the Law*, New York: Longman Publishing 1994.

Mulder, R.V. De & C. Van Noortwijk (1997), 'More science than art: Law in the 21st century', in: *Proceedings of the 12th Bileta Conference, The Future of Legal Education & Practice*, University of Durham, March 24th & 25th 1997, Part 5, p. 7-14.

Noortwijk C. van (1995), *Het woordgebruik meester* (Legal Word Use), with a summary in English, Dissertation, Lelystad: Vermande 1995, ISBN 90-5458-261-8.

Ulmer, S. (1967), 'Mathematical models for predicting judicial behaviour', in: Bernd, J.L. & Jones, A. (Eds.), *Mathematical applications in political science*, III, Charlottesville, p. 67.

Zeleznikow, J. & Hunter, D. (1994), *Building Intelligent Legal Information Systems*, Computer/Law series, part 13, Deventer/Boston: Kluwer 1994.