Comparing *n*-gram frequency distributions

Explorative research on the discriminative power of *n*-gram frequencies in newswire corpora

Thomas Bardoel

ANR: 306855

HAIT Master Thesis series nr. 12-012

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF ARTS IN COMMUNICATION AND INFORMATION SCIENCES,
AT THE SCHOOL OF HUMANITIES
OF TILBURG UNIVERSITY

Thesis committee:

Dr. M.M. van Zaanen
Prof. dr. E.O. Postma

Tilburg University
School of Humanities
Department of Communication and Information Sciences
Tilburg center for Cognition and Communication (TiCC)
Tilburg, The Netherlands
August 2012

Abstract

In the present study we have attempted to establish the viability of using *n*-gram frequencies to distinguish between parts of different newswire corpora. Each corpus was divided into two parts or subcorpora. For each value of *n* (1-8), each of the subcorpora was compared to both the other subcorpus originating from the same corpus (intra-corpus comparison) and to the subcorpora originating from different corpora (inter-corpus comparison). When comparing subcorpora, different rank thresholds were implemented, i.e. only the *r* highest ranked *n*-grams were considered. This rank threshold ranged from 100 to 1,000 with intervals of 100. Inspection of the *n*-gram frequency distributions revealed pieces of unremoved meta-text in several corpora. This considerably restricted the conclusions of this research. Results were most promising for $n = 1$ and higher rank thresholds and suggest the viability of this approach.

Acknowledgements

I remember Menno advising me and my fellow students at our first bachelor's thesis meeting in fall 2010 to start the actual writing as soon as possible, because it would benefit the process. Yet, many a time during the project of my master's thesis I found myself contemplating how to proceed without any result, only to finally decide to start typing and see where it would lead, resulting almost immediately in new ideas and insights. The point of this is that ultimately I have learned and will learn the things that need to be learned. It may, however, take me some time, which can put people's patience to the test. Gratitude for people's patience with me is therefore a recurring theme in these acknowledgements.

First of all, I want to thank my parents for their unconditional support, both morally and financially (the latter was in fact conditional, though very royal). Throughout my study they have always been encouraging and I am very appreciative of that. Because it has taken me slightly more time than planned to finish my study, I also want to thank them for their patience.

Furthermore I would like to thank Menno for supervising this thesis. He has provided me with extensive feedback and always took the time to discuss things with me. More than once I went into his office slightly cynical and perhaps even a little desperate about the way the research was proceeding. However, these meetings always led to new ideas and interesting viewpoints, turning my mood around. Because it has taken me slightly more time than planned to finish my master's thesis, I also want to thank him for his patience. In addition, I would like to thank Eric Postma for taking place in the thesis committee.

Finally I want to express my gratitude to my brother, my sister and my friends. For the past six years (and before) they have both inspired and entertained me and I am sure they will continue to do so in the future. Because I have a tendency to be late and I have kept them waiting on more than one occasion, I also want to thank them for their patience.

Thomas Bardoel Tilburg, August 2012

Table of contents

1 Introduction

Written texts can differ in many respects. Differences in writing styles can be based on various writer characteristics, like gender age, or descent (native vs non-native speakers) or text characteristics, like genre. Just as numerous causes of different writing styles can be distinguished, differences in writing styles can manifest themselves in many ways. Examples include vocabulary richness, frequency of function words, and sentence length (Zheng, Li, Chen, & Huang, 2006).

Writing styles can be investigated focusing on content, statistics, or both. For example, one could focus on content by investigating whether a writer uses a particular stylistic device in his texts. Focusing on both content and statistics, one could compare the texts of two writers, and examine whether one writer uses a particular stylistic device more frequently than the other does. Finally, one could focus on statistics, regardless of content. Zipf's law (Zipf, 1949) is in that respect a well known example. Zipf discovered that there is a pattern in the number of times specific word types occur in a text. Concretely his law states that word type frequencies relate to each other in the following way: the most frequent word in a text occurs twice as much as the second frequent word, three times as much as the third frequent word, etc.

Naturally, Zipf's law is an approximation, and there is not likely to be a natural language corpus which follows Zipf's law exactly. Nor will there be two corpora with all the same word frequencies. It could therefore be interesting to compare the word frequency distributions of two or more texts with each other, in order to find out if these distributions can be used to distinguish between different texts. Ideally then, the frequency distributions of two texts extracted from the same corpus would be relatively similar, whereas the frequency distributions of two texts extracted from different corpora would be relatively diverse.

Moreover, it would be interesting to explore the same possibility for word patterns. Perhaps the frequency distributions of patterns of a certain length tend to be more similar for related texts than for unrelated texts. Thus, the discriminative power of both word and pattern frequency distributions will be the topic of this thesis.

Aim of the research

In an exploratory study the difference in linguistic usage between several English written news sources will be investigated. Patterns of different sizes will be extracted from English written news corpora from both American and non-American news agencies. These patterns will then be arranged in order of occurrence. We want to find out whether the distributions of patterns differ between the different corpora. Moreover, portions of the same corpus will be compared to ensure that differences found between different corpora are meaningful. More concretely, an answer is sought to the following research question:

RQ: Is the distribution of patterns in corpora characteristic enough to distinguish between different corpora about the same topics?

Moreover, we want to find out whether patterns of a certain length are better suited for the task at hand than others. Thus the influence of different pattern lengths is explored in order to answer the following subquestion:

SQ1: How does pattern length influence the distinctive power of pattern distributions?

If the distribution of patterns resembles the distribution of individual words in a language, few instances will occur frequently, whereas most instances will not occur frequently. This can be visualized by a graph that starts very steep and has a long, almost horizontal tail. Because the differences in occurrence between instances in the tail of the distribution graph are expected to be very small, the distinctive power of pattern distributions, if present at all, will most likely reside in the left part of the distribution graph, with the most frequent instances. By ranking the instances in order of frequency (the most frequent pattern ranked 1, the second most frequent 2, etc.) we have the possibility to implement a rank threshold. This means that we restrict ourselves to the r highest ranked (i.e. most occurring) patterns. By changing the value of r we can then explore the following subquestion:

SQ2: How does rank threshold influence the distinctive power of pattern distributions?

The outline of this thesis is as follows. Chapter 2 provides the theoretical framework, starting with a discussion of the core notions of construction grammar. Next, the concept of *n*-grams is discussed, and it is explained how *n*-grams, which are essentially patterns of length *n*, relate to and can be seen as constructions. Chapter 2 is concluded with a discussion of Zipf's law and the extension of Zipf's law to *n*-grams. Subsequently, the experimental setup is described in chapter 3, starting with a description of the data, followed by an explanation of how the data are adapted and the patterns are extracted, and ending with a description of how the data are analyzed. The results section in chapter 4 provides the *n*-gram frequency distributions and the outcomes of the statistical tests. These results will then be discussed in chapter 5, and finally a conclusion is drawn in chapter 6.

2 Theoretical framework

2.1 Construction grammar

Construction grammar is a relatively new approach to modeling language, involving the broad notion of constructions. Simply put, constructions are form and meaning/function pairings and they include concepts such as "words, idioms, partially lexically filled and fully general linguistic patterns" (Goldberg, 2003, p. 219). Since constructions are pairings of form and meaning/function, they can be conceived of as signs in the Saussurean sense (Verhagen, 2009). Construction grammar allows for a complete account of our knowledge of language by means of what Goldberg calls a "construct-i-con", a network of constructions.

Whereas generative grammar theory considers semi-idiosyncratic constructions to be of less importance for linguistic theory, constructionist approaches do focus on these partially fixed constructions and advocate the possibility to extend the way of acquisition of these constructions to the regular core constructions of language (Goldberg, 2003). That is, both regular and irregular constructions are thought to be learned through induction, instead of the core principles of a language being innate, as generative approaches assume.

Schematicity

Constructions can be more or less schematic. An example of a relatively schematic construction is the transitive construction (S V O), in which a subject does something to an object. A less schematic construction would be the *hit*-construction, an instance of the transitive construction: S TO HIT O, in which the act that the subject performs on the object is known, namely hitting it. Even less schematic is the fully instantiated *John hits Paul*. This is in turn an instance of the *hit*-construction.

Another example of schematicity is the structure of compounds such as *tree hugger*, in which the second noun is the head and the first noun is the complement (a tree hugger is a kind of hugger and not a kind of tree). In turn, *hugger* is an instance of the more schematic *-er*-construction, in which a verb gets the suffix *-er* (whether or not with coercion) to result in a noun that describes an entity that performs the act described by the

verb.

Specificity and fixity continuum

Construction grammarians dismiss the idea of a clear distinction between lexicon and syntax (e.g. Verhagen, 2009; Taylor, 2002). A clear visualization of the way constructions relate to lexicon and syntax is provided by Doğruöz and Backus (2009). An adapted version of their specificity continuum is shown in figure 1. This specificity continuum depicts how constructions can range from (phonologically) fully specified to (phonologically) abstract.

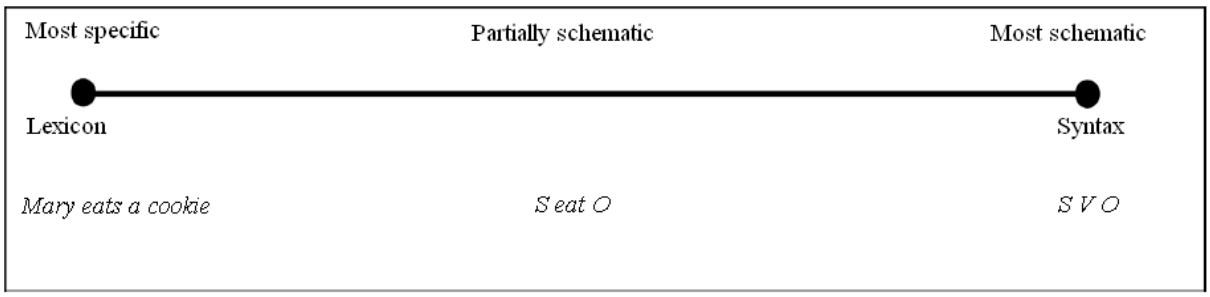


Figure 1. Specificity continuum adapted from Doğruöz and Backus (2009)

On the right side of the figure a schematic construction is shown, namely the transitive clause construction. This constructional schema has as one of its many instances the construction in the middle, in which a subject eats an object. This puts restrictions on the nature of both the subject and the object (e.g. the subject must be animate and the object must be edible). This *eat*-construction has in turn the fully specified construction on the left as one of its instances.

Analogous to the specificity continuum in Doğruöz and Backus (2009) a continuum of fixity can be visualized (figure 2).

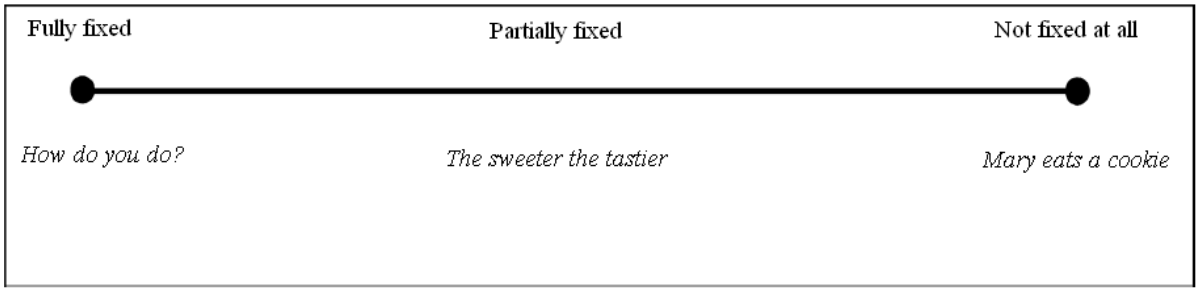


Figure 2. Fixity continuum

This continuum ranges from fully fixed (idiomatic) to not fixed at all (the regular/prototypical core of a language). The left side of the figure shows a construction that does not allow any variation (**How does Marcel do?*), whereas the meaning of the construction on the right side of the figure is a function of "general syntactic rules" (Taylor 2002, p. 575) and lexical elements (*Mary*, *eat* and *cookie*), which could be easily replaced by other ones, like *Roel*, *play* and *Tetris*. In between these two construction types are partially fixed constructions, like *The X-er the Y-er* (Fillmore, Kay, & O'Connor, 1988). Depending on their place on the fixity continuum these constructions can allow little (left) or much (right) variation.

2.2 *n*-grams

An *n*-gram is "an *n*-token sequence of words" (Jurafsky & Martin, 2009, p. 117). *n*-grams originate from the field of computational linguistics. They are used in probabilistic *n*-gram models to predict the next word after a sequence of $n - 1$ words. For example, the chance that the utterance *Brussels sprouts are* is followed by the word *disgusting* can be approximated by dividing the number of times the 4-gram *Brussels sprouts are disgusting* occurs in a text by the number of times the $(n - 1)$-gram *Brussels sprouts are* occurs.

n-grams are a much used concept in corpus research. Clough, Gaizauskas, Piao, and Wilks (2002) used, among other techniques, *n*-gram overlap to assess the reuse of newswire texts by newspapers. By viewing a newswire text and a newspaper text as sets of *n*-grams and comparing them they measured the proportion of *n*-grams shared by the texts, in order to determine to what extent the newswire text was reused. As opposed to Clough et al. (2002) we are not interested in the content of *n*-grams, but rather in their distribution over complete corpora from different sources.

Furthermore, *n*-grams are often used in authorship identification. For example, Argamon and Juola (2011) give an overview of the methods used by researchers competing in the international authorship identification competition at PAN 2011. Several of the participating researchers used *n*-grams as features for authorship attribution and authorship verification tasks. A distinction is made here between word *n*-grams and character *n*-grams. The former is a pattern of *n* words, whereas the latter is a pattern of *n* characters

In the present study we do not use *n*-grams for probabilistic models and we are not interested in their content. Rather, we simply use the concept of *n*-grams to state the size of patterns. An *n*-gram can thus be conceived of as a fully specified construction of length *n*.

We will investigate constructions ranging from one token, 1-grams or unigrams, to eight tokens, 8-grams. It should be noted that in this thesis a token does not necessarily correspond to a word. Punctuation marks like periods, commas and quotation marks are instances of tokens as well. So *the man, standing in the corner, waves* is a 9-gram, whereas *the man standing in the corner waves* is a 7-gram. Punctuation marks are included, since they add information to a construction. For example, the full stop in the 4-gram *to begin with* . indicates that the construction (3-gram) *to begin with* is placed at the end of a sentence, thus putting a restriction on its meaning. That is, a meaning in the form of (1) is possible, while (2) is ruled out.

(1) Your assumption is wrong to begin with.

(2) If you build a house, you have to begin with the fundament.

So, although we are going to look at *n*-gram frequencies regardless of their content, the reason for including punctuation marks does concern content.

2.3 Zipf's law

For a particular text the number of times each word occurs in it can be counted (term frequency). Subsequently the words can be ranked based on their frequency of occurrence, with the most common word ranked 1, the next most common word ranked 2, etc. Zipf (1949) described a regularity in the proportion between the rank *r* of a word and the actual frequency of occurrence of this word in a text. This regularity is known as Zipf's law, which mathematically states:

$$f = \frac{c}{r}$$

in which *f* is the frequency of a word in the text, *r* is its rank, and *c* is a constant. This means that the highest ranked word occurs twice as much as the second ranked word, which in turn occurs twice as much as the fourth ranked word, etc. Consequently, a graph depicting word frequency as a function of rank on a log-log scale will have a slope of -1. Figures 3a and 3b show the word frequency plotted as a function of rank for respectively

the Reuters-RCV1 corpus (Manning, Raghavan, & Schütze, 2008) and the Brown corpus of 1 million words of American English (Francis and Kucera, 1964, in: Ha , Hanna, Ming, & Smith, 2009) along with the curve predicted by Zipf's law. Ha et al. (2009) note that while studies following Zipf's discovery confirmed it for small corpora, for large corpora (with more than 1 million words) the curves start to become steeper from a certain rank. This tendency can be seen in figures 3a and 3b.

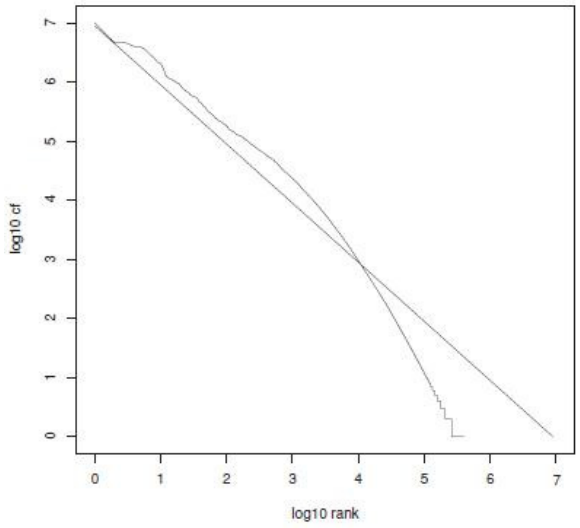


Figure 3a. (Manning et al., 2008)

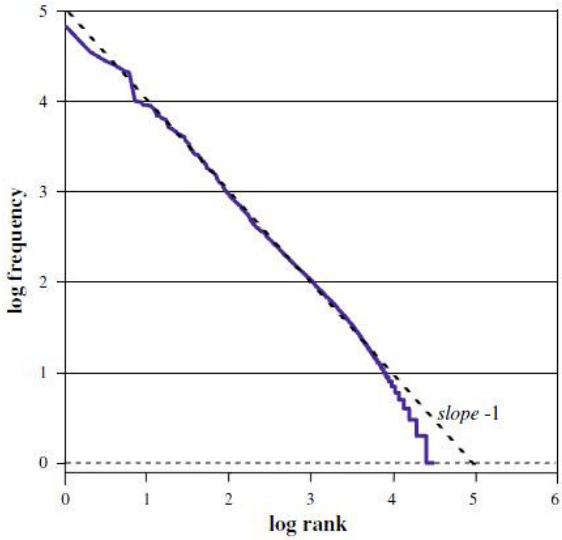


Figure 3b. (Francis & Kucera, 1964 in: Ha et al., 2009)

Ha, Sicilia-Garcia, Ming, & Smith (2003) looked at Zipf curves for *n*-grams ranging from $n = 2$ to $n = 5$ in large corpora. Figure 4 shows the frequency distribution of these *n*-grams, plus unigrams and the dotted curve predicted by Zipf's law, they found in the 19 million token Wall Street Journal 1987 corpus. Ha et al. (2003) report average slope decreases of the *n*-gram curves ranging from 0.66 ($n = 2$) to 0.59 ($n = 5$).

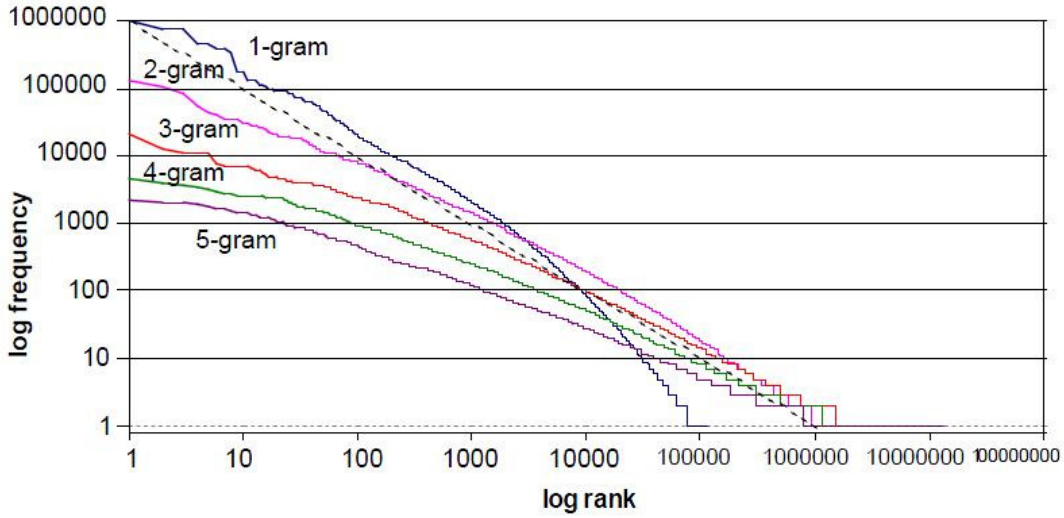


Figure 4. Zipf curves for the Wall Street Journal 1987 corpus (Ha et al., 2003)

As becomes clear from the foregoing, token (unigram) frequency distributions in large corpora seem to follow Zipf's law rather well up to a certain rank. Multigram ($n = 2$ and higher) frequency distributions, however, deviate from Zipf's law in two ways: they show gentle curves instead of straight lines and their average slope is smaller (absolute) than that of Zipf's curve.

In the present study we are going to compare the frequency distributions of *n*-grams (both unigrams and multigrams) of different large corpora. We are interested in finding out whether these distributions have distinctive power, and what influence the size of the *n*-grams and the number of ranks under consideration have on this potential distinctive power.

3 Experimental setup

3.1 Data

The corpora that are used in the present study come from the AQUAINT-2 Information Retrieval Text Research Collection (Voorhees & Graff, 2008). This collection contains the six English written newswire corpora shown in table 3.

corpus	abbreviation	tokens	country
Agence France-Presse	AFP	110.142.866	France
Associated Press	APW	82.015.295	USA
Central News Agency English Service	CNA	5.446.452	Taiwan
LA Times – Washington Post News Service	LTW	52.278.767	USA
New York Times	NYT	148.880.850	USA
Xinhua News Agency English Service	XIN	41.149.644	China

Table 3. Characteristics of the six corpora

The corpora consist of news articles from the time period October 2004 – March 2006. Because each corpus comprises the same period of time, mainly the same topics are expected to be covered. Therefore, probably roughly the same topic-specific words like names or dates occur in the different corpora. This decreases the chance of situation-specific words influencing the patterns, and therefore the pattern distribution. Moreover, the fact that presumably many of the same topics are covered on top of the fact that the compared corpora have the same genre, newswire texts, makes this endeavor a tough one. However, if *n*-gram distributions appear to be able to differentiate between corpora that are as much alike as the ones in this study, they can be powerful predictors in text classification tasks.

In order to be able to differentiate between different corpora based on *n*-gram distributions we need to make sure that distributions of different parts of the same corpus (intra-corpus pairs) do not vary as much as distributions of parts of different corpora (inter-corpus pairs). To that end all six corpora are split into two approximately equally large subcorpora (table 4). The first part of each subcorpus comprises the time period October 2004 – June 2005, while the second part comprises the period July 2005 – March 2006.

corpus	tokens	corpus	tokens	corpus	tokens
AFP1	57.571.100	CNA1	2.675.110	NYT1	72.610.520
AFP2	52.571.766	CNA2	2.771.342	NYT2	76.270.330
APW1	36.848.864	LTW1	25.645.887	XIN1	20.190.789
APW2	45.166.431	LTW2	26.632.880	XIN2	20.958.855

Table 4. Size of the twelve subcorpora

3.2 Finding patterns

To extract patterns from the raw data of the AQUAINT-2 Collection several steps need to be completed. First, a Perl program is implemented to remove all XML-code, leaving plain text. Next, the text is tokenized and sentensized, using Ucto-0.4.7, a tokenizer developed at Tilburg University (http://ilk.uvt.nl/downloads/pub/software/). The final step is the actual extraction of patterns from the tokenized corpora. This is done with Colibri (Van Gompel, 2011).

Colibri first encodes the text to a compressed binary form. It then invokes the pattern finder and finally decodes the binary code to a text file. This file contains, among other things, the patterns and their respective frequency of occurrence. Options are chosen so that $n = 1$ to $n = 8$ and the token frequency threshold = 10, which means that only n-grams that occur 10 times or more are retained. This is done in order to save space, since the amount of patterns occurring only once or a few times is expected to be very large. Considering the large size of the corpora, pruning n-grams occurring less than 10 times is unlikely to change the course of the distribution graph.

In addition, in order to control for effects of the varying sizes of the corpora the pattern counts are normalized. This is done by dividing the occurrence frequency of every pattern by the number of tokens of the corpus it is derived from (table 4).

3.3 Analysis

Because the more frequent n-grams presumably contain more information than the less frequent ones, a rank threshold will be implemented. For each n and for each corpus, the frequencies are ranked, with the most frequent n-gram having rank 1 (this is not necessarily the same n-gram for each corpus nor does it need to be, since we are comparing frequency distributions regardless of the content of the n-grams). When using

a rank threshold, we focus on the *r* most occurring *n*-grams, and disregard the rest. By implementing a rank threshold we thus focus on the left side of the distribution graph. This threshold is one of the two parameters that will have to be fine-tuned in order to find the optimal metric for differentiating between corpora based on *n*-gram distributions.

The other parameter that we will test is simply pattern length. By varying *n*, we can explore the distinctive power of distributions for patterns of different lengths. Examination of the interaction between rank threshold and pattern length will hopefully lead to a metric that can easily differentiate between the newswire corpora and can ideally be extrapolated to other corpora and genres.

Statistics

We want to compare the distributions of *n*-gram frequencies. To that end we employ the Mann-Whitney test. This is the non-parametric equivalent of the independent *t*-test (Field, 2009). Whereas the latter assumes a normal distribution, the Mann-Whitney test does not. This is essential, as the Shapiro-Wilk test shows that the *n*-gram distributions of all (sub)corpora are significantly different from a normal distribution ($p < .001$).

While the Mann-Whitney test can be used to check for differences between two independent samples, the Kruskal-Wallis test can be used to test for differences between more than two independent groups. It is the non-parametric equivalent of the ANOVA, and therefore does not assume normality either. A post hoc test for Kruskal-Wallis can be conducted by doing multiple Mann-Whitney tests. Normally, this method requires making an adjustment to the critical value for significance, namely dividing it by the number of tests you conduct. This technique is called a Bonferroni correction (Field, 2009). For example, if post hoc procedures were to be conducted for all combinations of the twelve subcorpora (66 combinations), a critical value of .05 would be reduced to .05 / 66 = .0076.

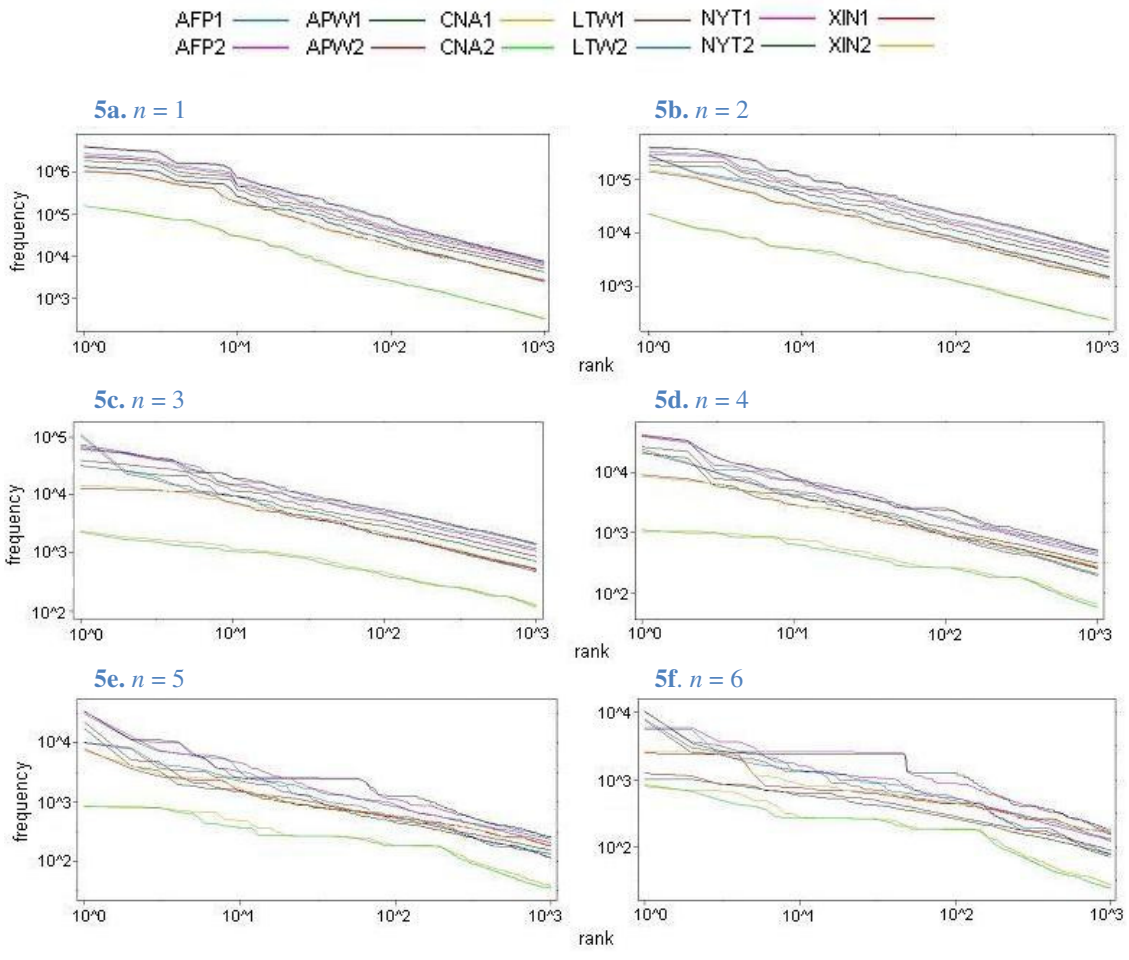
However, the Bonferroni correction is not relevant to our endeavor. We are not looking for hypothesized differences between (sub)corpora based on a pre-established critical value for significance. Rather, we just want the *p*-values to differ in a way that enables us to classify parts of the same corpus as the same and parts of different corpora as different, regardless of a significance level. Ideally though, *p*-values of tests between

parts of different corpora are as low as possible, whereas for tests between parts of the same corpus they are as high as possible. Because there is no need for a predefined critical value, we can just perform Mann-Whitney tests on all combinations of subcorpora without making any adjustments.

4 Results

4.1 *n*-gram frequency distributions

Frequency distributions of *n*-grams in the twelve subcorpora are displayed in figure 5. These frequencies have not yet been normalized. This makes it easier to see the distributions of the different corpora and it reflects the size of the corpora. With normalized frequencies the distribution graphs would be closer together, thus making it harder to distinguish between them. The graphs depict frequencies of the 1,000 highest ranked *n*-grams of each subcorpus for *n*=1 to *n*=8.



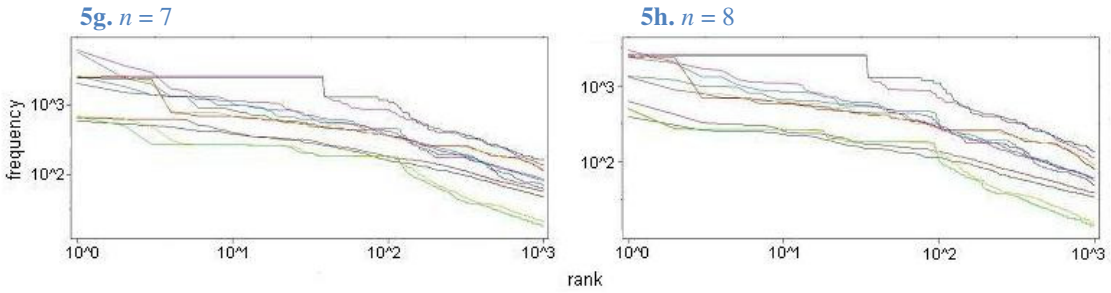


Figure 5. Frequency distributions of *n*-grams (*n* = 1 to *n* = 8) in subcorpora; non-normalized, rank threshold *r* = 1,000.

The distribution graphs for *n* = 1, *n* = 2 and *n* = 3 follow a rather straight course. From *n* = 4 up, however, graphs of several subcorpora start to contain increasingly large horizontal parts. These horizontal parts represent multiple *n*-grams with exactly or nearly the same frequency. As the value of *n* increases these horizontal parts shift to the left and upwards. The increasing size of the horizontal parts is a result of their shifting position and the logarithmic scale of the figure. That is, because of the logarithmic scale a graph part on the left of the x-axis is larger than a part on the right side depicting the same number of *n*-grams.

To examine the tendency described above we can for example focus on the distribution of *n*-grams in the NYT2 subcorpus. The NYT2 distribution graph for *n* = 4 contains a horizontal part on the right side of the x-axis. This part depicts 50 4-grams, namely the ones ranging from *r* = 56 (freq = 2,465) to *r* = 105 (freq = 2,440). In the 5-gram distribution graph of the NYT2 subcorpus the horizontal part has become larger and moved to around the middle of the x-axis. This part depicts 46 5-grams, namely the 5-grams from *r* = 13 (freq = 2,464) to *r* = 58 (freq = 2,440). In the distribution graph for *n* = 6 the horizontal part has again increased in size and shifted to the left and upwards, depicting 42 6-grams, ranging from *r* = 5 (freq = 2,464) to *r* = 46 (freq = 2,440). For *n* = 7 this tendency still continues, with the horizontal part depicting 38 7-grams, ranging from *r* = 1 (freq = 2,464) to *r* = 38 (freq = 2,464). Finally, at *n* = 8, the horizontal part depicts 34 8-grams, ranging from *r* = 1 (freq = 2,464) to *r* = 34 (freq = 2,460).

This pattern is summarized in table 5. If we check the frequencies of 3-grams we can identify a range that probably contains the ones that are also present in the horizontal graph parts of higher *n*-gram distributions. These are the 3-grams ranging from approximately *r* = 408 (freq = 2,465) to *r* = 454 (freq = 2,440). This is, however, difficult

to spot from the graph, because the part of the graph representing these 3-grams is very small due to the logarithmic axes. The 2-gram and 1-gram distribution graphs do not contain these deviating graph parts, since the ranks of these *n*-grams do not exceed the rank threshold of 1,000.

n	**begin *r***	**(freq)**	**end *r***	**(freq)**	**size**
3	408	(2,465)	454	(2,440)	47
4	56	(2,465)	105	(2,440)	50
5	13	(2,464)	58	(2,440)	46
6	5	(2,464)	46	(2,440)	42
7	1	(2,464)	38	(2,440)	38
8	1	(2,464)	34	(2,440)	34

Table 5. Characteristics of the horizontal parts in the NYT2 *n*-gram distribution graphs.

The graph deviations described in table 5 are highly unlikely to be coincidental and are presumably the result of unremoved meta-text in the NYT2 subcorpus. After inspection of the tokenized NYT2 subcorpus it indeed appeared to contain some meta-text that had not been filtered out during the preparing of the corpus. Based on the *n*-grams that occurred exactly 2,464 times, the two sentences in box 1 could be detected. These sentences also occurred 2,464 times in the corpus.

> All clients receive all budgets , but only full-service clients receive all stories .
> Please check your level of service to determine which stories you will receive .

Box 1. Unremoved meta-text from the NYT corpus.

Table 6 shows for each *n* the number of unique *n*-grams that the meta-text (MT) in box 1 contains (the second row). Furthermore it shows how many of these *n*-grams occur more than 2,464 times in the NYT2 subcorpus, i.e. the *n*-grams that also occur outside the meta-text (third row). From $n = 5$ up this number is 0, which means that all *n*-grams from $n = 5$ to $n = 8$ are exclusively found in the meta-text. The fourth row shows the average frequency of the meta-text *n*-grams in the NYT2 subcorpus. The large difference in average frequency between unigrams and multigrams is a result of punctuation marks and function words in the meta-text.

n	1	2	3	4	5	6	7	8
unique *n*-grams in MT	22	24	23	22	20	18	16	14
MT *n*-grams with frequency > 2,464	22	22	11	2	0	0	0	0
avg. frequency of MT *n*-grams in NYT2	518,265	8,148	2,618	2,464.1	2,464	2,464	2,464	2,464

Table 6. Characteristics of the *n*-grams in the meta-text in box 1.

As the value of *n* increases, the proportion of *n*-grams with a higher frequency than the meta-text decreases relative to the number of unique *n*-grams that can be extracted from the meta-text. This is a result of the fact that an (*n*+1)-gram can by definition not occur more often than an *n*-gram, since the first includes the latter. That is, if the 3-gram *All clients receive* occurs 2,464 times, the 4-gram *All clients receive all* must also occur 2,464 times (because of the meta-text occurring 2,464 times). Yet it is possible, that a 3-gram, say *your level of*, occurs more than 2,464 times (freq = 2,476), but a 4-gram (*check your level of*) occurs 2,464 times. Therefore, the proportion of (*n*+1)-grams that occur both in- and outside the meta-text relative to the number of unique (*n*+1)-grams that the meta-text contains can never be larger than the proportion of *n*-grams.

The occurrence of unremoved meta-text explains both the presence and the shifting of the horizontal distribution graph parts. Whereas in natural language the frequency of a certain *n*-gram typically decreases as *n* increases, the frequency of an *n*-gram extracted from a piece of meta-text will never drop below the frequency of the meta-text (in the case of box 1 2,464). Therefore *n*-grams extracted from meta-text (especially if it occurs relatively often as in the case of box 1) shift to the left as *n* increases and eventually become the most frequent *n*-grams, when all natural *n*-gram frequencies drop below the meta-text frequency.

4.2 Comparing subcorpora

Intra-corpus comparisons

Mann-Whitney testing on intra-corpus pairs is done in order find out to what extent the two subcorpora of each corpus differ from each other. The results of these analyses are shown in figure 6. *p*-values of the Mann-Whitney test are compared to the various rank thresholds employed when performing the test. These rank thresholds range from 100 to

1,000, with intervals of 100.

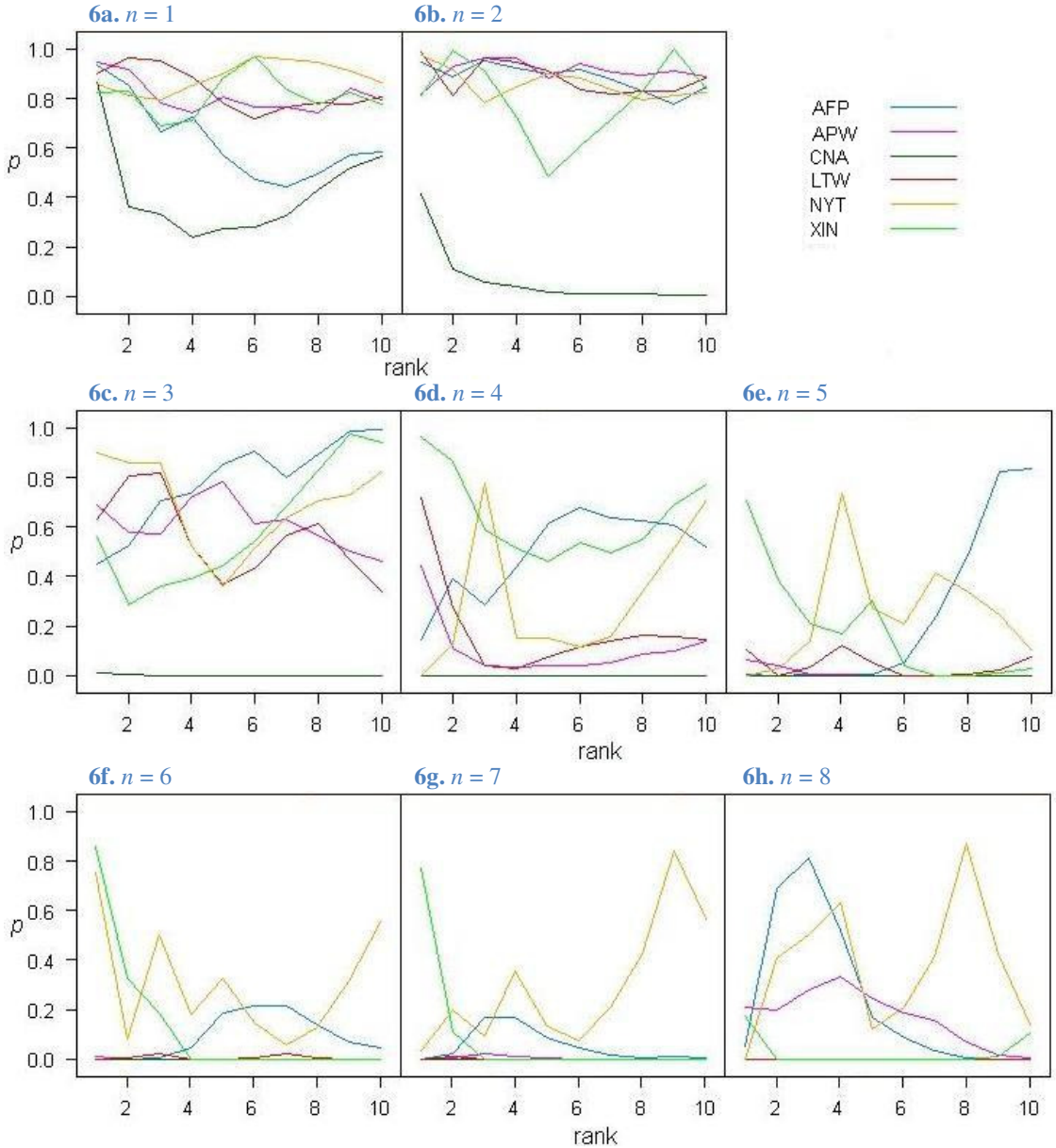


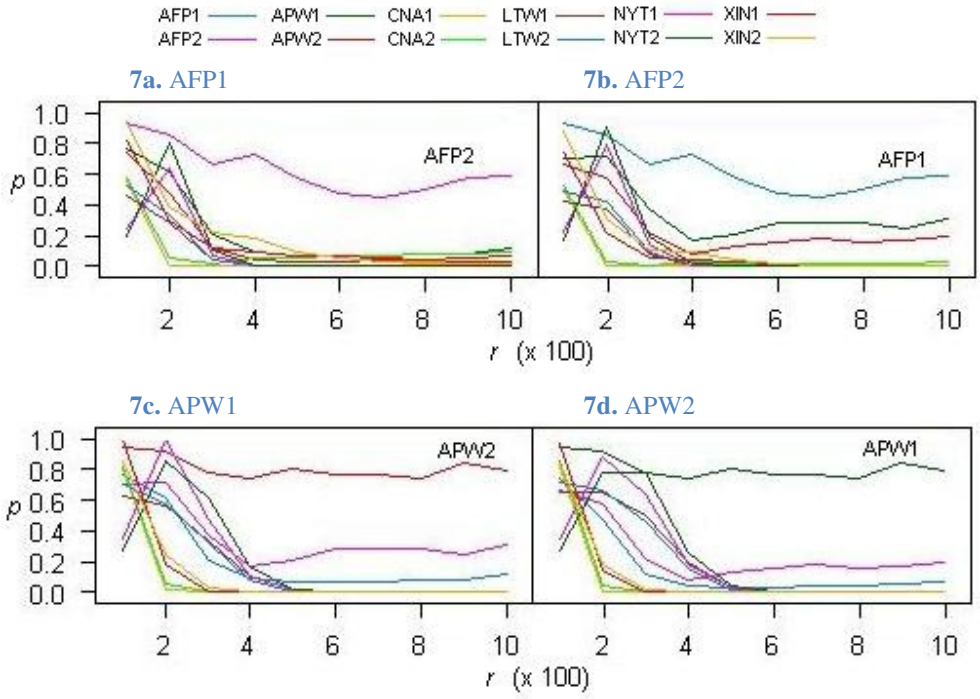
Figure 6. Intra-corpus Mann-Whitney testing (n = 1 to n = 8); Mann Whitney p-value is plotted as a function of rank threshold (r); rank threshold ranges from 100 to 1,000 with intervals of 100.

With the exception of the CNA corpus, distributions of low n-grams (n = 1, n = 2, n = 3) seem to be rather similar for subcorpora of the same corpus, with $p > .4$ (n = 1, n = 2) and $p > .3$ (n = 3) for all rank thresholds. Distributions of n-grams of n = 4 and higher (figures 6d-h) tend to show larger differences between intra-corpus pairs, with p-values frequently approaching 0. As a result, we cannot use n-gram distributions of n = 4 and higher to distinguish between corpora. After all, if n-gram distributions differ greatly between parts

of the same corpus, differences found between parts of different corpora are meaningless. The fact that higher *n*-gram distributions differ more between intra-corpus pairs than lower *n*-gram distributions is probably caused to a large extent by the unremoved meta-text in the corpora. In lower *n*-gram distributions the meta-text has little or no effect, but, as described above, this effect increases as the value of *n* increases. We therefore restrict ourselves to comparing the distribution graphs of inter-corpus pairs for $n = 1$ (figure 7), $n = 2$ (figure 8), and $n = 3$ (figure 9), on which the meta-text has little or no influence.

Inter-corpus comparisons

Mann-Whitney testing on inter-corpus pairs is done in order find out to what extent the parts of different corpora differ from each other. Figure 7 shows the results of Mann-Whitney testing on all possible pairs of subcorpora for $n = 1$. *p*-values are again a function of the different rank thresholds (100-1000, with intervals of 100).



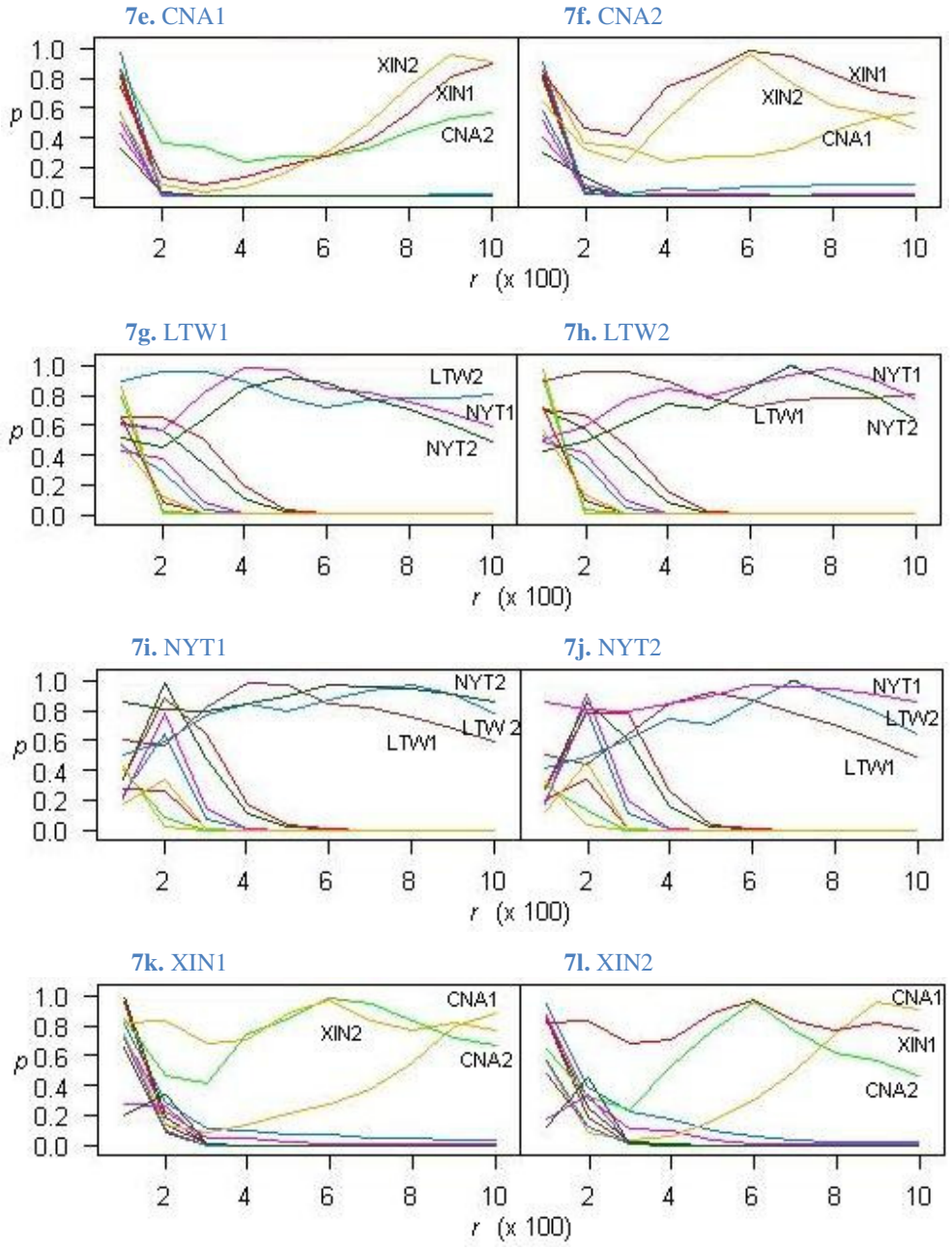
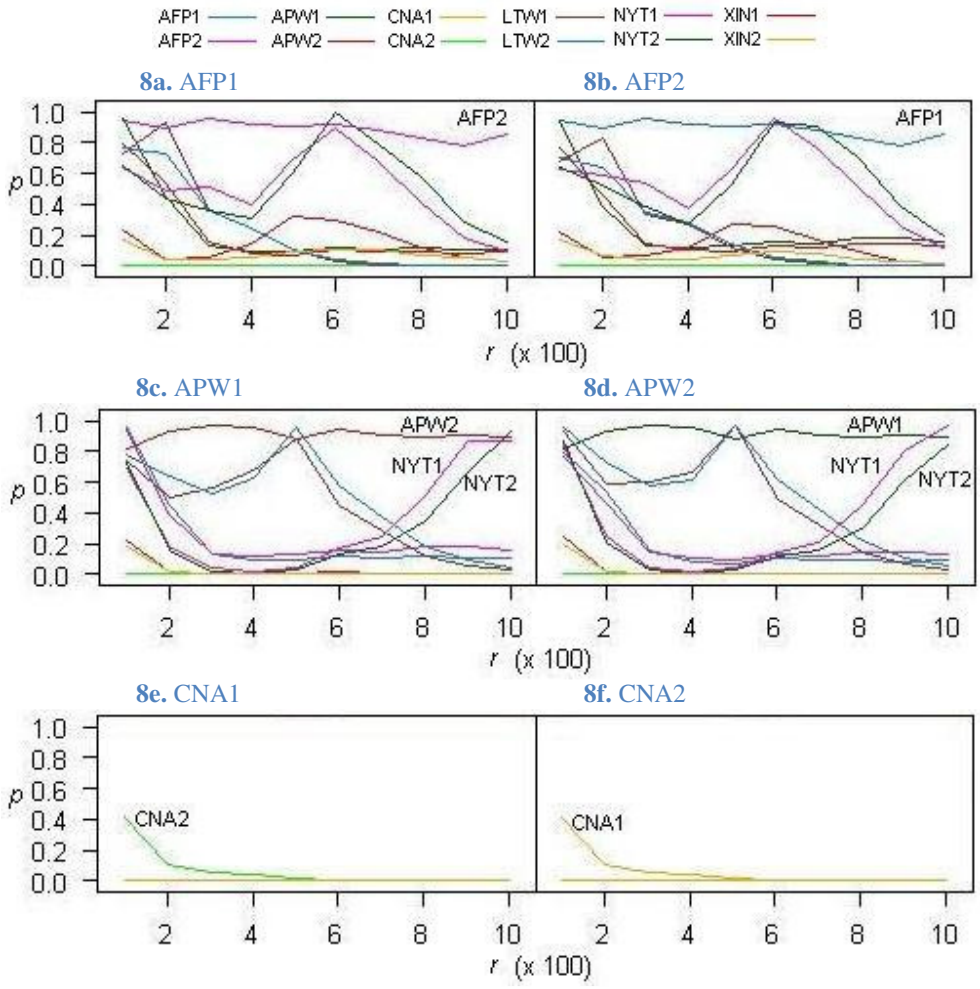


Figure 7. Inter-corpus Mann-Whitney testing (n = 1); Mann Whitney p-value is plotted as a function of rank threshold (r); rank threshold ranges from 100 to 1,000 with intervals of 100.

At the left side of the figures, at low rank thresholds, 1-gram distributions seem to have no distinctive power, because many inter-corpus pairs have high p-values. This changes as the rank threshold increases. The increasing rank threshold causes many subcorpora to have low p-values and only one or few subcorpora to have high p-values. Examination of graphs at higher rank thresholds reveals the pairing of the following subcorpora:

- AFP1 and AFP2
- APW1 and APW2
- CNA1, CNA2, XIN1, and XIN2
- LTW1, LTW2, NYT1, and NYT2

Figure 8 shows the results of Mann-Whitney testing on all possible pairs of subcorpora for $n = 2$. p-values are again a function of the different rank thresholds (100-1000, with intervals of 100).



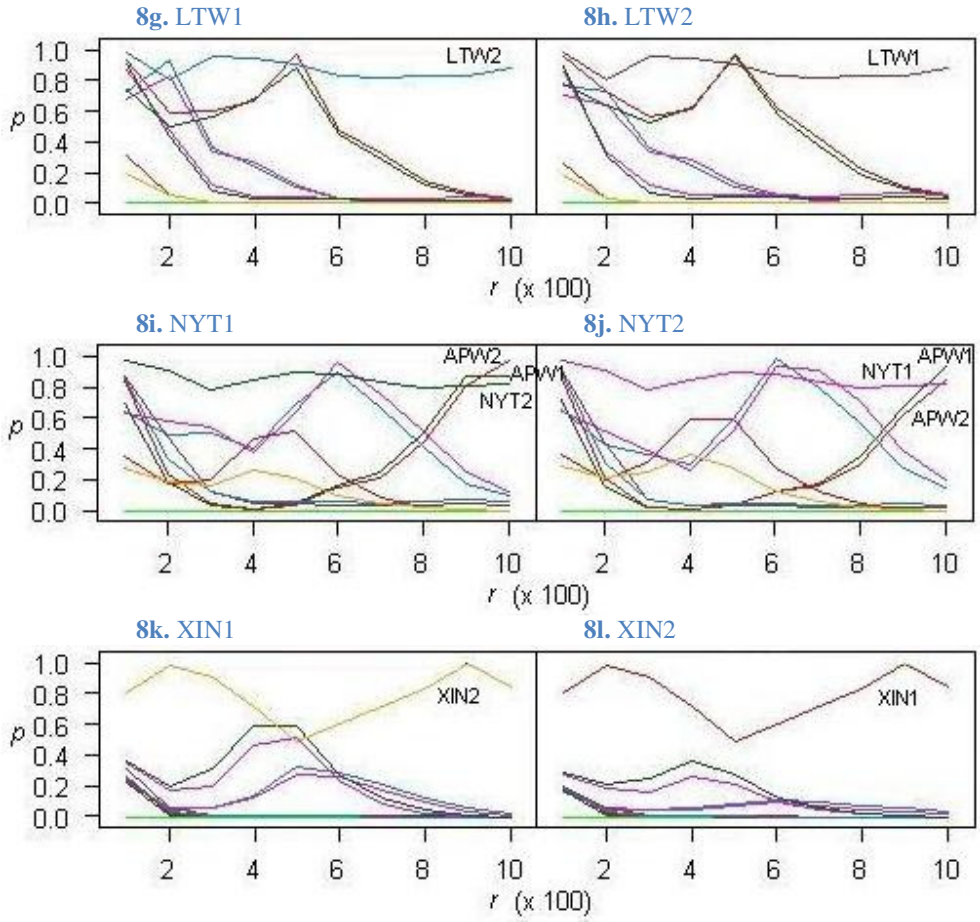


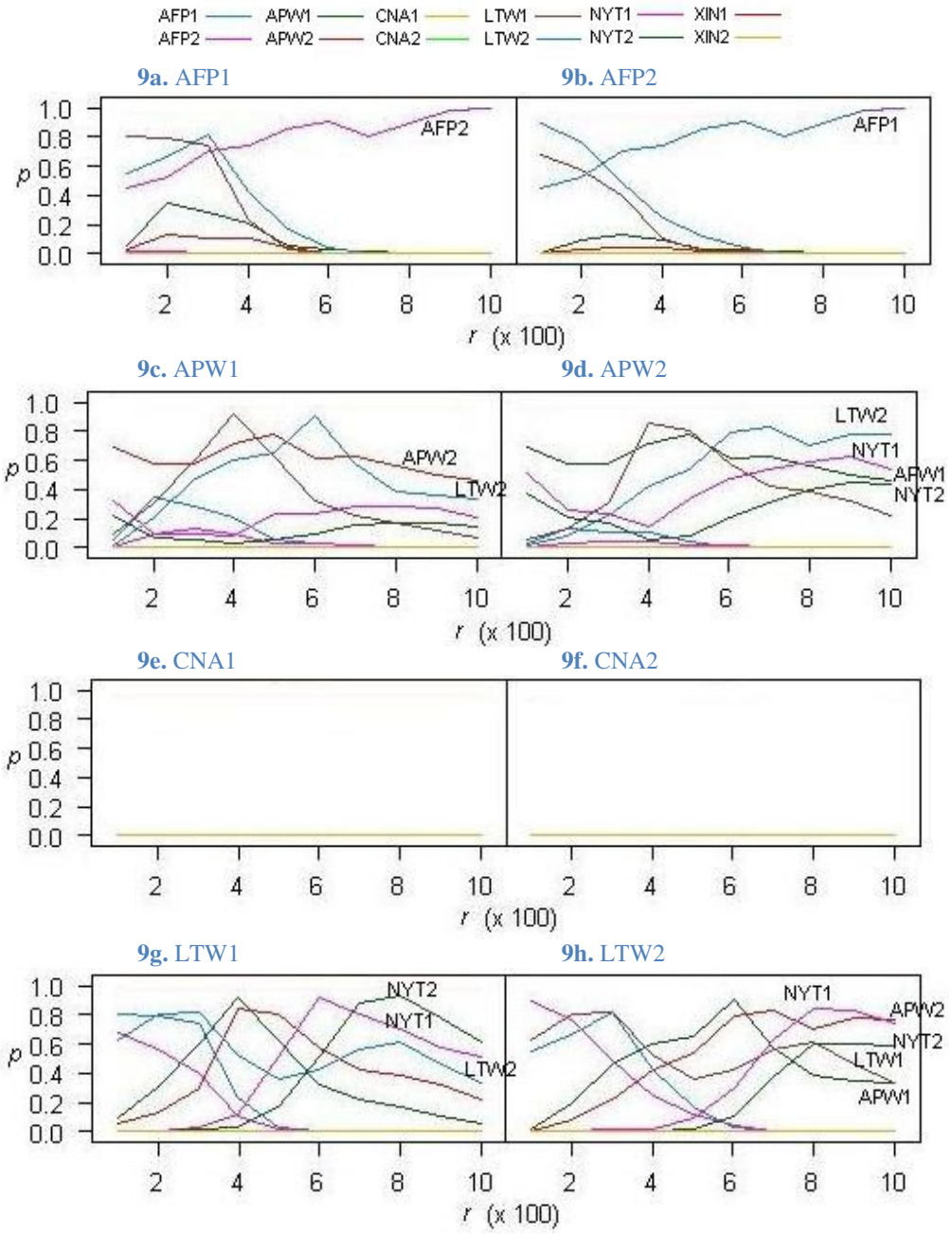
Figure 8. Inter-corpus Mann-Whitney testing ($n = 2$); Mann Whitney p-value is plotted as a function of rank threshold (r); rank threshold ranges from 100 to 1,000 with intervals of 100.

At $n = 2$, again, the frequency distributions seem to have no distinctive power when low rank thresholds are applied, due to the many pairs with high p-values. Distinctions become easier to make as r increases, but in general the graphs follow a less steady course at higher rank thresholds than the unigram graphs in figure 7. Graphs of intra-corpus pairs, however, still follow a relatively steady course. Examination of graphs at higher rank thresholds reveals the pairing of the following subcorpora:

- AFP1 and AFP2
- APW1, APW2, NYT1, and NYT2
- LTW 1 and LTW 2
- XIN 1 and XIN2

For the CNA1 and CNA2 subcorpora all inter-corpus comparisons result in p-values barely higher than 0, which is why only two lines can be seen in figure 8e and 8f.

Figure 9 shows the results of Mann-Whitney testing on all possible pairs of subcorpora for n = 3. p-values are again a function of the different rank thresholds (100-1000, with intervals of 100).



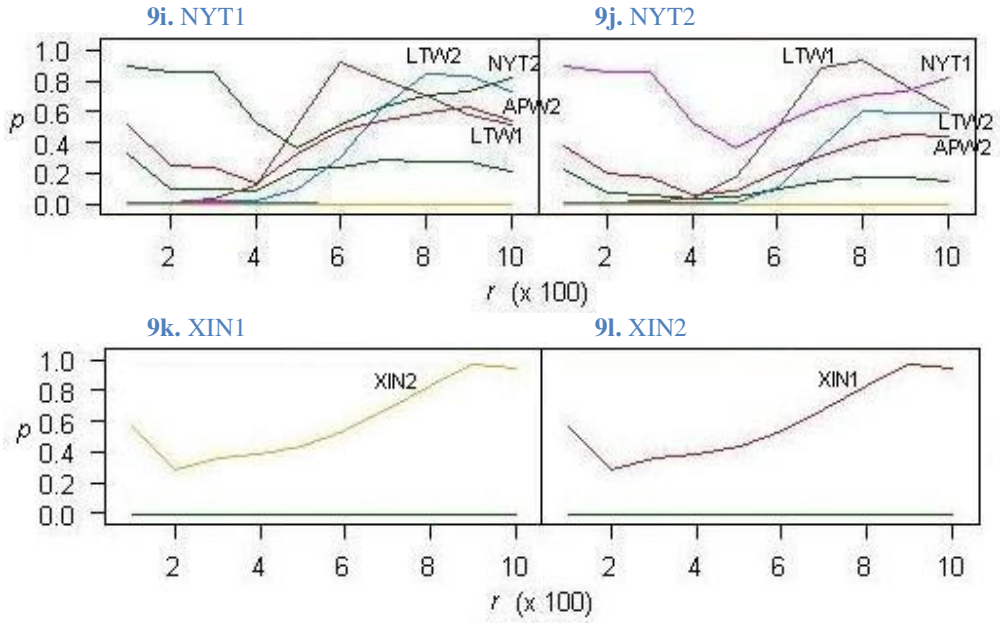


Figure 9. Inter-corpus Mann-Whitney testing (n = 3); Mann Whitney p-value is plotted as a function of rank threshold (r); rank threshold ranges from 100 to 1,000 with intervals of 100.

At n = 3 distinctions between some subcorpora are becoming more blurry. This can perhaps be ascribed to the effects of the meta-text which begin to show. We should therefore view these results with caution. It is, however, still possible, to group the following subcorpora according to their p-values at higher rank thresholds:

- AFP1 and AFP2
- APW1, APW2, LTW1, LTW2, NYT1, and NYT2
- XIN1 and XIN2

Again, inter-corpus comparisons with CNA1 and CNA2 result in p-values barely higher than 0. Moreover, now even the intra-corpus comparisons with CNA1 and CNA2 result in p-values very close to 0, as has been shown earlier in figure 6. Therefore only one line can be seen in figures 9e and 9f.

The outcomes summarized in figeres 7, 8, and 9 are conspicuous if we view them in the light of the origin of the corpora (table 3, p.10). Subcorpora of the only French corpus (AFP) can in all three cases only be grouped together. The same holds for subcorpora of the Chinese corpus (XIN), which can also only be grouped together, except for when n = 1, in which case distributions of parts of the XIN corpus also resemble distributions of parts of the Taiwanese CNA corpus. Finally, the three American corpora

can be grouped together in at least one of the three cases, while they can never be grouped with non-American corpora.

In all cases subcorpora of the same corpus are grouped together. This is an outcome we were expecting and hoping to achieve. However, it should be taken into account that there are considerable differences in size between inter-corpus pairs of subcorpora, whereas the differences in size between intra-corpus pairs are marginal. Although we tried to control for the effects of these differences by normalizing the pattern frequencies, we cannot be certain that the size differences do not play a role in this at all.

5 Discussion

Flaws in the data have had a significant influence on the outcomes of our endeavor to distinguish between different corpora based on *n*-gram frequency distributions. Several corpora appeared to contain some unanticipated passages of meta-text, which influenced these distributions increasingly as the value of *n* became higher. As a result *n*-gram frequency distributions of $n = 4$ and higher could not be meaningfully compared.

As for the lower *n*-grams that have been analyzed, some interesting things have come to light. For example, with respect to the second subquestion of this study, as to how rank threshold influences the distinctive power of *n*-gram distributions, it appears that in the range of $r = 100$ to $r = 1{,}000$ the higher thresholds seem to benefit the distinctive power of the frequency distributions. Furthermore, especially at $n = 1$ and $n = 2$, most of the subcorpora seem to be relatively similar to subcorpora which can be expected to be similar based on origin both in terms of "mother corpus" (i.e. similarity between intra-corpus pairs) as in terms of country (e.g. similarity between American corpora).

As a result of the flawed data, the first subquestion, as to how *n* influences the distinctive power of *n*-gram distributions, is a difficult one to answer. However, if we look at the Mann-Whitney graphs for 1-grams (figure 7) and for 2-grams (figure 8) one striking difference is that the 1-gram graphs follow a much more steady course than the 2-gram graphs. That is, although at both $n = 1$ and $n = 2$ clear distinctions can be made between similar and dissimilar subcorpora at a high rank threshold ($r = 1{,}000$), only for $n = 1$ can the same clear distinctions be made at middle rank thresholds (around $r = 400$ to $r = 600$). Because of their steady course, 1-gram distributions seem more reliable than 2-gram distributions.

Since the present study has an explorative nature, many methodological choices can be adjusted for further exploration. These include:

-*omission of punctuation marks*; in the present study punctuation marks are included as tokens, since they add information to a construction. However, in several studies observing *n*-gram frequency curves punctuation marks are excluded (e.g. Ha et al., 2009; Ha et al., 2003)

-*different languages*; the corpora in this study all consist of English texts. Although differences between corpora from English-speaking countries and non-English speaking countries are highlighted, it might be even more interesting to investigate differences and similarities between different languages.

-*creating intra-corpus pairs of different size*; the intra-corpus pairs in this study are of approximately the same size, whereas inter-corpus pairs differ in size. Although we tried to compensate for these differences by normalizing the *n*-gram frequencies, it might be appropriate to create the same circumstances for both intra-corpus and inter-corpus comparisons, including a difference in size that has to be dealt with.

-*a different normalization technique*; in this study, normalization of *n*-gram frequencies is done by dividing them by the number of tokens of their respective corpus. An alternative possibility is to extract the number of *n*-grams from each corpus and divide *n*-gram frequencies by that number. This means that the divisor for normalizing 1-grams remains the number of tokens of a (sub)corpus, but for every higher *n*-gram the divisor decreases with the number of sentences of the (sub)corpus (since every sentence contains one (*n*+1)-gram less than it contains *n*-grams). Seeing that the variable under consideration is *n*-gram frequency, it might be more appropriate to normalize based on number of *n*-grams, and thus use different normalizations for each *n*, rather than on number of tokens (1-grams) for every *n*.

-*extend rank threshold*; in the present study a maximum rank threshold of 1,000 is implemented. It appeared that differences between corpora could be better observed at higher rank thresholds than at lower ones, so it might be worthwhile to inspect differences at rank thresholds higher than 1,000.

-*inclusion of skipgrams*; skipgrams are *n*-grams with one or more unspecified tokens or "wildcards". The inclusion of skipgrams makes it possible to consider not only fully specified constructions, as is the case in this study, but also less specified ones. For example, a partially lexically filled construction like *twistin' the night away* (Jackendoff, 1997) can be captured by the skip-5-gram *He *** the *** away*.

6 Conclusion

In the present study we have attempted to establish the viability of using *n*-gram frequencies to distinguish between parts of different corpora. The idea behind this approach was that *n*-grams are essentially fully specified constructions of length *n*, and that *n*-gram frequency distributions are thus a representation of the use of constructions by a source. Six newswire corpora of different sizes and from different countries (although all written in English) were used for this purpose. Each corpus was divided into two parts or subcorpora. *n*-grams of $n = 1$ to $n = 8$ were extracted from these subcorpora and ranked based on their frequency (the most frequent *n*-gram having rank 1, the second most frequent rank 2, etc.).

For each value of *n*, each of the subcorpora was then compared to both the other subcorpus originating from the same corpus (intra-corpus comparison) and to the subcorpora originating from different corpora (inter-corpus comparison). Comparisons were made by means of Mann-Whitney testing. Ideally intra-corpus comparisons would lead to high *p*-values and inter-corpus pairs would lead to low *p*-values.

When comparing subcorpora, different rank thresholds were implemented, i.e. only the *r* highest ranked *n*-grams were considered. A rank threshold was implemented based on the assumption that frequency differences in higher ranks are larger than in lower ranks, and that higher ranks therefore contain more distinctive power than lower ranks. This rank threshold ranged from 100 to 1,000 with intervals of 100.

Inspection of the *n*-gram frequency distributions revealed pieces of unremoved meta-text in several corpora. This considerably restricted the conclusions of this research, since intra-corpus comparisons at $n = 4$ and higher showed that subcorpora of the same corpus were not alike, which was probably caused to a large extent by the flawed data.

However, under the circumstances in the present study we can conclude the following with respect to the research question:

Is the distribution of n-grams in corpora characteristic enough to distinguish between different corpora about the same topics?

This exploration of the distinctive power of *n*-gram frequency distributions has neither

confirmed nor rejected the definite viability of this approach. However, some findings suggest that this approach might have something to offer in text classification. In particular the comparing of 1-gram distributions seems promising, since a clear distinction can be made between similar and dissimilar subcorpora, which is in addition relatively constant over longer ranges of rank thresholds. Moreover, all intra-corpus pairs are relatively similar and inter-corpus pairs that are relatively similar originate from the same country or in one case from countries with the same language (China and Taiwan).

The latter is also true for 2-grams and 3-grams, but here the distinctions are less clear and less constant over longer ranges of rank thresholds. Comparisons at these values of *n* might benefit from higher rank thresholds. Extending the rank threshold might cause distinctions to be clearer and more constant.

As for the *n*-grams of $n = 4$ and higher, no valid conclusions can be drawn, since it is unclear to what extent the unremoved meta-text has contributed to the dissimilarity of intra-corpus pairs. Although the influence of the meta-text is probably considerable, we cannot be sure that intra-corpus pairs would be similar with clean data. Future research on this topic might clarify this.

References

Argamon, S., & Juola, P. (2011). Overview of the international authorship identification competition at PAN-2011. *Proceedings of CLEF 2011, PAN competition notebook.* Retrieved from http://www.webis.de/research/events/pan-11

Clough, P., Gaizauskas, R., Piao, S. S. L., & Wilks, Y. (2002). METER: MEasuring TExt Reuse. *Annual Meeting of the ACL.* Retrieved from http://portal.acm.org/citation.cfm?id=1073083.1073110

Doğruöz, A. S., & Backus, A. (2009). Innovative constructions in Dutch Turkish: An assessment of on-going contact induced change. *Bilingualism: Language and Cognition, 12*(1), 41-63.

Field, A. (2009) *Discovering statistics using SPSS* (3d ed.). London: Sage Publications Ltd.

Fillmore, C. J., Kay, P., & O'Connor, M. C. (1988). Regularity and idiomaticity in grammatical constructions. *Language, 64*, 501-538.

Goldberg, A. E. (2003). Constructions: a new theoretical approach to language. *Trends in cognitive sciences, 7*(5), 219-224.

Ha, L. Q., Sicilia-Garcia, E. I., Ming, J., & Smith, F. J. (2003). Extension of Zipf's law to word and character *N*-grams for English and Chinese. *Computational Linguistics and Chinese Language Processing, 8*(1), 77-102.

Ha, L. Q., Hanna, P., Ming, J., & Smith, F. J. (2009). Extending Zipf's law to *n*-grams for large corpora. *Artificial Intelligence Review, 32*, 101-113.

Jackendoff, R. (1997). Twistin' the night away. *Language, 73*, 534-559.

Jurafsky, D., & Martin, J. H. (2009). *Speech and Language Processing*. Upper Saddle River, NJ: Pearson Education.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval.* Cambridge, England: Cambridge University Press.

Taylor, J. R. (2002). *Cognitive Grammar*. New York, NY: Oxford University Press.

Verhagen, A. (2009). The conception of constructions as complex signs. Emergence of structure and reduction to usage. *Constructions and frames, 1*, 119-152.

Zheng, R., Li, J., Chen, H., & Huang, Z. (2006). A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology, 57*(3), 378-393.

Zipf, G. K. (1949). *Human behaviour and the principle of least effort.* Reading, MA: Addison-Wesley.

Equipment

Gompel van, M. (2011). *Colibri*. Radboud University Nijmegen. Retrieved from https://github.com/proycon/colibri

Voorhees, E., & Graff, D. (2008). *AQUAINT-2 Information-Retrieval Text Research Collection.* Linguistic Data Consortium, Philadelphia.