N-grams and the clustering of registers

Stefan Th. Gries[1], John Newman[2], and Cyrus Shaoul[2]
University of California, Santa Barbara[1] and University of Alberta, Edmonton[2]

Abstract
This paper explores the use of different-length *n*-grams as a basis for identifying relationships between registers. Using the BNC Baby and the ICE-GB corpora as test cases, we study the following questions in a principled rigorously bottom-up fashion: (i) how well can *n*-grams distinguish parts in corpora? (ii) how much do corpus parts derived from *n*-grams correspond to registers or sub-registers as defined by corpus compilers? (iii) which *n*-gram length yields the best discriminatory power? (iv) how many *n*-grams yield the best discriminatory power? In contrast with most previous studies, we use hierarchical agglomerative cluster analyses to cluster sub-corpora and use average silhouette widths as a diagnostic to decide on how many clusters to distinguish.[*]

Key words
register, spoken vs. written, *n*-grams, BNC Baby, ICE-GB, hierarchical agglomerative cluster analysis, (average) silhouette width

1. Introduction

One of the perennially hot topics in corpus linguistics is the study of different registers.[1] While the field of register studies is quite diverse, it is possible to roughly categorize them from several different perspectives. One perspective is concerned with what constitutes the focus of the study. On the one hand, a study can focus on the distribution of some linguistic feature(s) and register only plays a role in as much as it allows one to account for, or partial out, variability in the distribution of said linguistic feature(s). Thus, in this case, the dependent variable, the 'thing' to be explained, so to speak, is the linguistic feature. This focus would lead to statements such as "Linguistic feature *F* is rather frequent in language as it occurs *x* times per million words, but it is worth pointing that *F* is not distributed uniformly across registers: it is much more frequent in register *R* than in register *S*." On the other hand, the study can focus on register *R* and *S* as such, and the different linguistic feature(s) whose distribution is explored are just different ways in, or dimensions with respect to, which *R* and *S* are described and contrasted. This type of orientation would lead to statements such as "*R* is more interactive than *S*, given that features *F*, *G*, and *H* are significantly more frequently than in S." Crucially, the linguistic features involved in both types of approaches can include many very different characteristics. For example Biber's multidimensional analysis involves morphological, syntactic, and lexical descriptors (cf. Biber 1988 for an overview and below for more references), but other studies have also looked at character *n*-grams (cf. Cavnar & Trenkle 1994), key words (cf. Xiao & McEnery 2005), 2-grams (Crossley & Louwerse 2007), 3-grams (Orasan & Krishnamurthy 2002), 4-grams (Biber, Conrad, & Cortes 2004).

Yet another perspective relates to the direction of the research: top-down or bottom-up. Most studies of register are done in a top-down manner, meaning that they are based on, for

instance, register divisions implemented in (hand-compiled, hand-categorized) corpora. Obviously, such studies assume register is a relevant factor in the sense that different registers exhibit distributions of linguistic variables that are, following Occam's razor, sufficiently different from each other to merit making the distinction in the first place. This line of research is closely intertwined with the topics of corpus homogeneity and corpus granularity (cf. Kilgarriff 2001, Gries 2006). Other studies, by contrast use a bottom-up approach to determine the degree with which particular registers may actually qualify as text types, given their marked linguistic differences. The best-known studies adopting this perspective are again those involving Biber's multidimensional approach to registers and their linguistic correlates (cf. Biber, Conrad, & Cortes 2004; Biber, Csomay, Jones, & Keck 2004; Csomay & Cortes 2010; but also cf. Santini 2007; Mota 2010 as well as Xiao & McEnery 2005 and Nishina 2007 for overviews).

The present paper is largely exploratory in nature, but exhibits characteristics of several of the above approaches. While Biber's multidimensional approach is doubtlessly the most sophisticated approach to date, it also involves much supervised, semi-manual, if not completely manual, annotation, which makes it difficult to apply to large or constantly changing corpora. While alternative descriptors have been explored (see the work cited above), there are few studies that explore several different descriptors (with differing levels of resolution) or different settings of descriptors and at the same time validate the results against an independently arrived-at gold standard.

In this paper, we address several of these issues. Springing from Biber et al.'s (1999:990-1024) observation that features of *n*-grams can distinguish registers, we will consider the question of if and how well different kinds of lexical *n*-grams can distinguish between the registers that corpus compilers have used to guide their decisions of what to include in a corpus. Compared to many linguistic features that have been used in the multidimensional approach, *n*-gram frequencies are easy to compute and so if we were to obtain accurate classifications with these *n*-gram frequencies, this would provide evidence that *n*-gram measures are a valid computational shortcut to arrive at homogeneous sub-corpora from which text could be sampled for further study of phenomena that are sensitive to the differing lexical distributions.

At the same time, we also explore the question of which *n*-grams are best suited for this task, and how many should be included to obtain the largest degree of discriminatory power. In the studies cited above the analysis often involves studying *n*-grams that contain either three or four words (3-grams or 4-grams) whose frequency exceeds a particular threshold (e.g., 10 or 15 occurrences per million), but it is unclear whether larger or smaller *n*-grams would not result in a more accurate discrimination between registers (compared to some gold standard). It is also unclear which frequency threshold will yield the best results.

Given the above descriptions of current practice, it could be said that our study is top-down in the sense that as we rely on the register classifications carried out by creators of two well-known corpora of British English as our gold standard. At the same time, it is also bottom-up in the sense that we extract *n*-grams in a completely unsupervised manner, we do not take an arbitrary size of *n*-gram *a priori*, and we do not assume *a priori* that a particular number of *n*-grams is the best set to base our diagnostic. To recap, we address the following four questions:

i. How well can *n*-grams distinguish parts in corpora?
ii. How much do corpus parts based on *n*-grams correspond to registers or sub-registers as defined by corpus compilers?
iii. Which *n*-gram length yields the best discriminatory power?
iv. How many *n*-grams yield the best discriminatory power?

This kind of approach has several advantages over previous work on register. This work explores a computationally low-cost alternative to Biber's multidimensional approach in a similar vein to Xiao & McEnery's (2005) key words approach. What does our approach offer?

- it does not require a reference corpus for comparison;
- it avoids their risky strategy of using a British English corpus as a reference corpus for American English data;
- it does not a priori single out one particular *n*-gram length but is more corpus-driven in that it explores differently long *n*-grams, which for the longer *n*-grams begins to include syntactically revealing *n*-grams;
- explores different levels of corpus granularity since more than one division of the corpus is considered: we do not just consider different corpus parts on the level of the register *or* the sub-register, but compare results for both;
- it uses exploratory statistical methods widely used in Information Retrieval and Text Classification, namely hierarchical agglomerative cluster analysis, plus an additional statistic to be outlined below.

In the following section, we will outline how we studied the above questions in detail.

2. Methodology

In this study, we explore *n*-gram distributions and cluster discriminability in two different corpora: the British National Corpus Baby (<http://www.natcorp.ox.ac.uk>) and the British Component of the International Corpus of English (<http://www.ucl.ac.uk/english-usage/projects/ice-gb/>). Both corpora exhibit considerable internal structure, which is represented in Table 1 and Table 2 respectively.

Corpus register	Sub-register
academic	applied sciences vs. arts vs. belief/thought vs. natural sciences vs. social sciences vs. world affairs
demographic spoken	AB vs. C1 vs.C2 vs. DE
fiction	imaginative
news	applied sciences vs. arts vs. belief/thought vs. commerce vs. leisure vs. natural sciences vs. social sciences vs. world affairs

Table 1: Our classification of the parts of the BNC Baby

Corpus register	Sub-register
spoken dialog	private vs. public
spoken monolog	scripted vs. unscripted
spoken mix	broadcast
written printed	academic vs. creative vs. instructional vs. non-academic vs. persuasive vs. reportage
written non-printed	letters vs. non-professional

Table 2: Our classification of the parts of the ICE-GB

For each corpus, we first generated and stored one frequency list of all case-insensitive sentence-internal *n*-grams excluding numbers, punctuation marks, and other 'special' characters.[2] For this retrieval and counting process, *n* was 5≥*n*≥1 and "sentence-internal" refers to the corpus compilers' notion of *sentence*, which for the spoken data can also correspond to utterances and/or turns; the frequencies compiled were *log* frequency of occurrence + 1. Each table was then sorted according to the *n*-grams overall frequency in the corpus and divided into 100 slices of 1 to *x* percent of the *n*-grams, thus starting with the most frequent 1% of all *n*-grams, then the most frequent 2% of all *n*-grams, etc. up to the last slice containing 100%, i.e. all, *n*-grams (cf. Chujo 2004 and Nishina 2007 for similar approaches).

Second, for each corpus we computed hierarchical agglomerative cluster analyses (HCAs) on each of the 100 *n*-gram frequency list slices. We used the Pearson measure shown in (1) as the measure of similarity and Ward's method as our amalgamation rule.

$$(1) \qquad \sum_{i=1}^{n} (freq_{part1} freq_{part2}) \div \sqrt{\sum_{i=1}^{n} freq_{part1}^{2} \times \sum_{i=1}^{n} freq_{part2}^{2}}$$

Third, to measure the quality of each clustering produced by the HCA, we calculated a measure of cluster quality called the silhouette width.[3] The average silhouette width is defined as the mean of the silhouette widths for all possible clustering solutions, and we will use the maximal average silhouette width (MASW) as our measure of the quality of a clustering solution. For each of the 100 cluster analyses per *n*-gram per level of corpus granularity we computed the average silhouette widths and determined which number of clusters returned the MASW. For instance, we did one cluster analysis on the 19 sub-registers of the BNC Baby based on the 1% most frequent 1-grams. Since this means we clustered 19 sub-registers, there were 17 theoretically possible numbers of clusters one might assume: 2, 3, 4, …, 16, 17, and 18 – assuming 1 or 19 clusters does not make sense since it amounts to saying the data are all completely homogeneous (as all the sub-registers form one cluster) or completely heterogeneous (as no sub-registers forms a cluster with any other one). We then computed for the average silhouette width for each of these 17 numbers of clusters, and then chose the number of clusters *n* with the MASW, which is the one characterized by the largest amount of discrimination between sub-registers. Then, we repeated this with the 2% most frequent 1-grams, then the 3% most frequent 1-grams, etc., until the 99% most frequent 1-grams, and all 1-grams. Then, the whole process was also done for 2-grams, 3-grams, 4-grams, and 5-grams.

This approach – the nesting of computing average silhouette widths for different numbers of clusters within cluster analyses for 100 different slices within *n*-gram lengths within a particular corpus granularity – is graphically represented in Table 3, in which they grey-shaded cell corresponds to the computation described in the examples above.

For our evaluation, the critical aspect then was whether the cluster solutions at the MASW correlated with the gold standard registers and sub-registers or not. If *n*-grams (of particular *n*-gram sizes and particular percentage slices) were an appropriate way to cluster (sub-)registers, then the cluster solutions with the largest average silhouette widths should correlate with (sub-)registers in the corpus. If, on the other hand, the cluster solutions were completely at odds with the gold standard division of the corpus into (sub-)registers, then *n*-gram frequency might not be a very good heuristic.[4] One example is shown in Figure 1, which contains the results of the analysis for the grey-shaded cell in Table 3. The left panel shows the percentage slices of 1-grams on the *x*-axis (from 1 to 100), the average silhouette widths on the *y*-axis, and the numbers plotted into the coordinate system indicate the number of clusters that

yielded the MASW. The right panel shows the dendrogram for the MASW solution. That is, when the 19 sub-registers of the BNC Baby are clustered on the basis of 1-grams, then the cluster solution with the highest discriminatory power arises already when only the 1% most frequent 1-grams are included.

Corpus	Granularity	*N*-gram	1% slice	2% slice	…	99% slice	100% slice
BNC Baby	4 registers	1	HCA + 2 ASWs, → 1 MASW	HCA + 2 ASWs, → 1 MASW	…	HCA + 2 ASWs, → 1 MASW	HCA + 2 ASWs, → 1 MASW
		2	HCA + 2 ASWs, → 1 MASW	HCA + 2 ASWs, → 1 MASW	…	HCA + 2 ASWs, → 1 MASW	HCA + 2 ASWs, → 1 MASW
		3	…	…	…	…	…
		4	…	…	…	…	…
		5	…	…	…	…	…
	19 sub-registers	1	HCA + 17 ASWs, → 1 MASW	HCA + 17 ASWs, → 1 MASW	…	HCA + 17 ASWs, → 1 MASW	HCA + 17 ASWs, → 1 MASW
		2	…	…	…	…	…
		…	…	…	…	…	…
ICE-GB	5 registers	1	HCA + 3 ASWs, → 1 MASW	HCA + 3 ASWs, → 1 MASW	…	HCA + 3 ASWs, → 1 MASW	HCA + 3 ASWs, → 1 MASW
		…	…	…	…	…	…

Table 3: Summary of the methodology
(HCA = hierarchical cluster analysis, (M)ASW = (max.) average silhouette width)

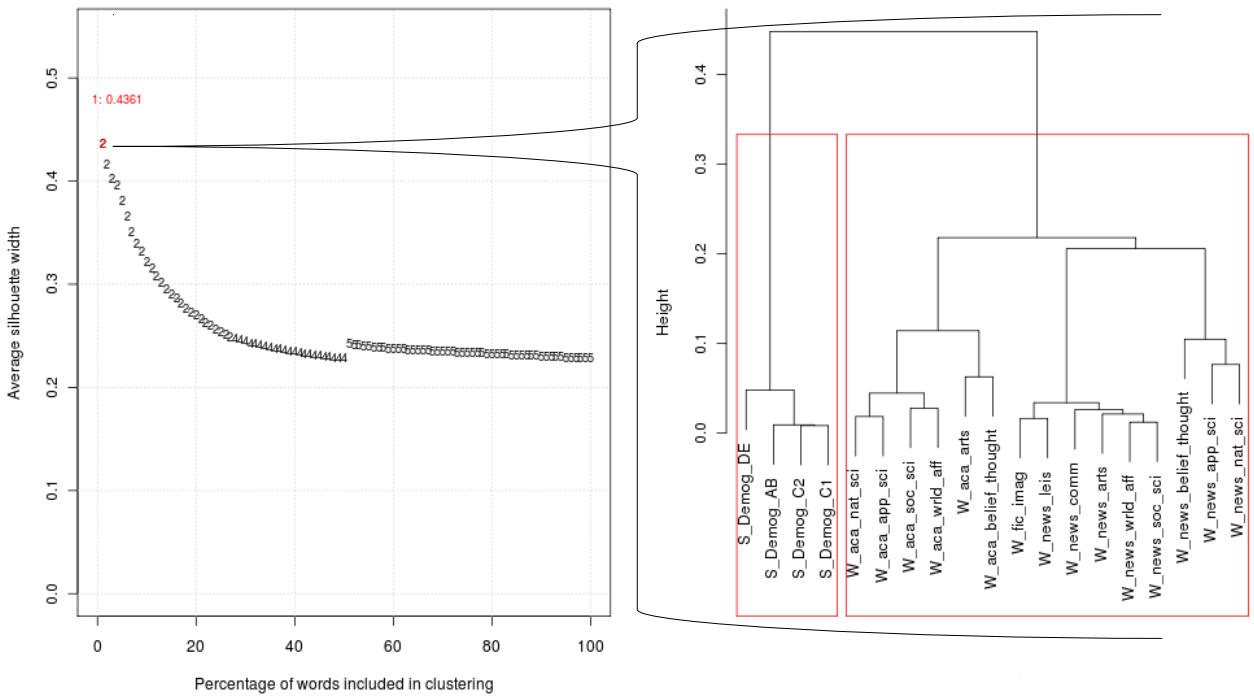

Figure 1: Number of clusters resulting from 100 cluster analysis on 1-grams in the 19 sub-registers of the BNC Baby (on the basis of MASWs; left panel) and the dendrogram resulting from the cluster analysis with the largest MASW for all 1-grams (on the basis of the 1%-slice (right panel)

The MASW suggests a 2-cluster solution that perfectly distinguishes spoken from written data. It also does a fairly good job at keeping the academic writing parts as well as the news parts together and even has some nice substructures such that, e.g., the humanities in academic are together (arts and belief/thought) or the imaginative fiction is in one cluster with leisure news.

Now that we have described our analytical methods, we will discuss the results of our cluster analyses.

3. Results

In this section, we will report the results of our cluster analyses for each of level of the corpus and each level of register granularity. In each sub-section, we will report three types of information:

- at *which %-slice increment* for each type of *n*-gram we found the MASW. This is to determine how many of the most frequent *n*-grams result in the highest discriminatory power, and in the above example in Figure 1, the answer would be "1."
- *how many clusters* were returned at the MASW in most of the cluster analyses (or, if the latter results are very varied, at all 10%-slice increments). This is to determine how much structure the cluster analyses find in the *n*-grams of the corpus parts. For instance in the above clustering shown in Figure 1, the answer would be "Including the most frequent 27% 1-grams yielded two clusters, including between 28% and 50% of the most frequent 1-grams returned four clusters, including 51% and more yielded five clusters."
- what the *structures of these clusters* are. This is to determine how well the *n*-gram-based clusters correspond to the corpus compilers' gold standard, and in the above example, the answer would be "The 2-cluster solutions (shown in Figure 1) divided the corpus into spoken vs. written ([spk] vs. [wrt]); the 4-cluster solutions divided the data up into [spk] vs. [acad news:socsci (with news:socsci correctly clustered with acad:socsci)] vs. [news:sci news:bel/tho] vs. [news:rest fiction]; the 5-cluster solutions that is nearly identical to the four-cluster solution but with the main difference that the cluster [acad] is split up into one containing natural and applied science writings and one that contains the remaining academic writing registers."

3.1 The BNC Baby and its 4 broad registers

Since there are only four elements to be clustered and, thus, only cluster numbers of two and three are possible, it is not surprising that there is little variation in the number of clusters across all *n*-gram lengths and across all %-slices. In fact, the numbers of clusters returned is always two. However, there is some variation as to what are included in the clusters:

- for 1-grams, the best corpus separation (i.e., the highest MASW) is found when the top 7% of the most frequent 1-grams are included, and including up to approximately the top 20% of all 1-grams returns [spk] vs. [acad fic news] clusters (i.e., [spk] vs. [wrt]); including more than that returns [acad] vs. [spk fic news] (i.e. the rest of the corpus);
- for 2-grams, the best corpus separation is found when the top 2% of the most frequent 2-grams are included ([spk fic] vs. [acad news]), including up to approximately the top 10% of all 2-grams returns [spk] vs. [wrt] clusters; including around 20% returns [acad] vs. [fic news spk], more than that returns [spk fic] vs. [acad news];

- for 3-grams, 4-grams, and 5-grams, the best corpus separation is found with the top 1% of the most frequent *n*-grams, and the clusterings return [spk fic] vs. [acad news].

Intuitively, the overwhelming trend for the 2-cluster solution [spk fic] [aca news] makes sense – there could be much worse/confusing solutions, e.g. [spk aca] [fic] [news]. Upon further consideration, it appears that 1-grams are somewhat problematic: the dendrograms are far from stable, the majority solution from all other *n*-grams is never returned, and their MASW results are rather erratic, as is indicated in Figure 2 below.

On the other hand, the 1-grams return [spk] vs. [wrt] and are, thus, maybe only more useful for more coarse-grained distinctions. However, these are only tentative conclusions since the coarse resolution of only four broad registers does not really allow conclusive validation of the results, which is why we now turn to the more fine-grained resolution of 19 sub-registers.

3.2 The BNC Baby and its 19 sub-registers

With regard to the best corpus separation, the results are very unanimous: the highest MASW is without exception found with already only the 1% most frequent *n*-grams. With regard to the number of clusters and their make-up, the results are more heterogeneous:

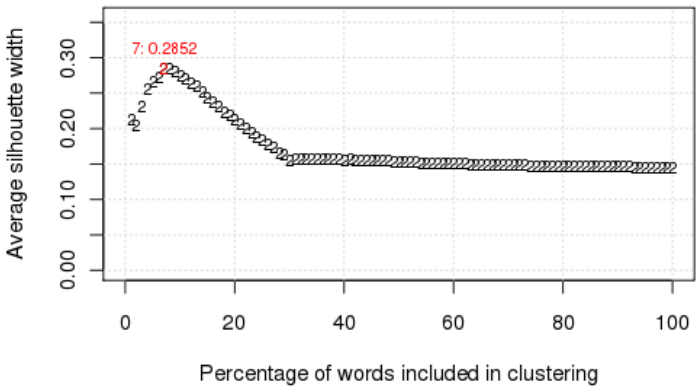

Figure 2: Number of clusters resulting from 100 cluster analysis on 1-grams in the four registers of the BNC Baby (on the basis of MASWs)

- for 1-grams, the 27 most frequent %-slices return [spk] vs. [wrt]; 28% to 50% return the four clusters [spk] vs. [acad] vs. [news:sci] vs [news:other fic], and including more than that returns [spk] vs. [acad:hardsci] vs. [acad:softsci] vs. [news:sci] vs. [news:other fic];
- for 2-grams, 3-grams, and 4-grams, nearly all cluster analyses return [spk] vs [wrt];
- for 5-grams, nearly all of the 17 most frequent %-slices return nearly exactly the above 4-cluster solution for 1-grams: [spk] vs. [acad] vs. [news:sci] vs [news:other fic]; including more than that returns 3 clusters [spk] vs. [$news_1$] vs. [fic $news_2$ {aca}].

This shows that spoken language is always recognized as different from the rest, but it also shows academic writing is often clustered separately and that sometimes at least there is even useful and meaningful structure within the clusters, as when academic sub-disciplines yield meaningful groups within clusters.

3.3 The BNC Baby: interim summary

The previous two sections have revealed several fairly clear trends, both in terms of content and

methodology. With regard to the former, two kinds of cluster solutions are particularly frequent: the 2-cluster solution [spk] vs. [wrt] and the 4-cluster solution [spk] vs. [acad] vs. [$news_1$ fic] vs. [$news_2$], both of which show that the distinction of spoken vs. written is a valid one here and that academic writing is very different from other writing; in addition, if anything, fiction is close to news.

With regard to the latter, the cluster solutions at the MASWs make good sense and *n*-grams have appreciable discriminatory power: the cluster solutions, while not all completely homogeneous, are rather good even if their frequency is not compared to a reference corpus and even if only a very small percentage of the most frequent *n*-gram types is included. The one kind of *n*-gram for which this is only partially true is 1-gram, which exhibits a large degree of volatility.

3.4 The ICE-GB and its 5 broad registers

For this corpus and that level of granularity, the best degrees of corpus separation are again obtained very early, i.e. with very small %-slices: for 1-grams, the top 4% of the *n*-grams yield the most structure, but for all other *n*-grams, only the top 1% is needed. With regards to the number of clusters, however, there is some variation despite the fact that five registers only allow for three possible clusters. We found one consistent result: our cluster analyses always distinguish [spk:broadcast] from the rest of the corpus, but there was considerable variation in how the rest of the corpus is clustered. At the MASW, 1-grams lump all other corpus parts together (and only produce better results when more than 60% of all 1-grams are included), while all other *n*-grams split the rest of the corpus up into [spk:other] vs [wrt], which is much more useful. However, there are also some dendrograms (for 4-grams and 5-grams with lower average silhouette widths) that split the two written registers up, which is again not particularly revealing.

3.5 The ICE-GB and its 13 sub-registers

With regard to the best corpus separation, the results were unanimous: the highest MASW is without exception found by using only the frequencies of the most frequent 1% of the *n*-grams. With regard to the number of clusters and their make-up, the results are much more heterogeneous:

- for 1-grams, the cluster solution at the MASW returns two clusters, which is represented in Figure 3. Interestingly, the division is not a simple [spk] vs. [wrt], but two of the written sub-registers are grouped into one larger cluster with the spoken data. Fortunately (for the method), these two written sub-registers are the two ones that, if any, one would want to group with [spk], namely [wrt:creative] and [wrt:letters].
  Also, interestingly, the more of the 1-grams that are included, the more clusters are returned: the 3-cluster solution creates a new cluster [spk:broadcast wrt:reportage wrt:persuasive], which is noteworthy because at least the first two are arguably related registers. The 4-cluster solutions with even more 1-grams are very similar but make [wrt:instructional] its own cluster.
- the other *n*-grams yield various different cluster solutions, whose numbers of clusters vary between 2 and 10. While this sounds as if the results were erratic, they were quite reasonable. The two by far most frequent cluster solutions for all *n*>1-grams are the 2-cluster solution from Figure 3 and, as the most frequent solution, the 4-cluster solution represented in Figure 4.

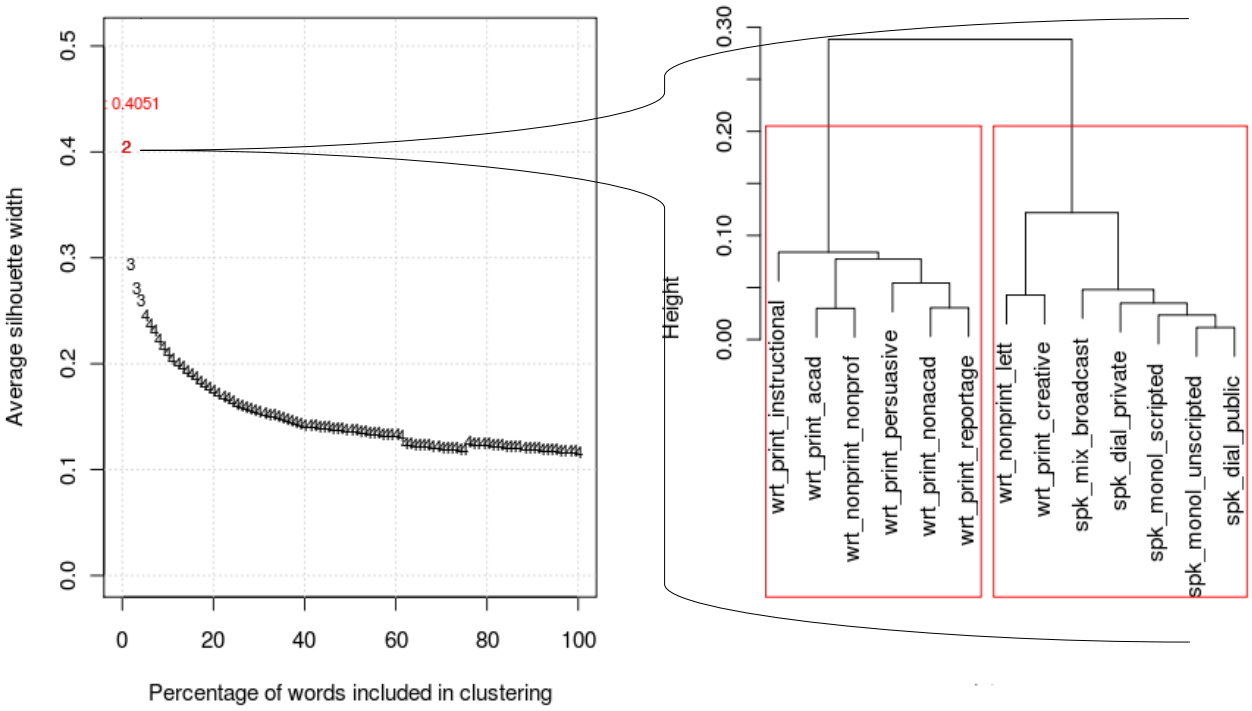

Figure 3: Number of clusters resulting from 100 cluster analysis on 1-grams in the 13 sub-registers of the ICE-GB (on the basis of MASWs; left panel) and the dendrogram resulting from the cluster analysis with the largest MASW for all 1-grams (on the basis of the 1%-slice (right panel)

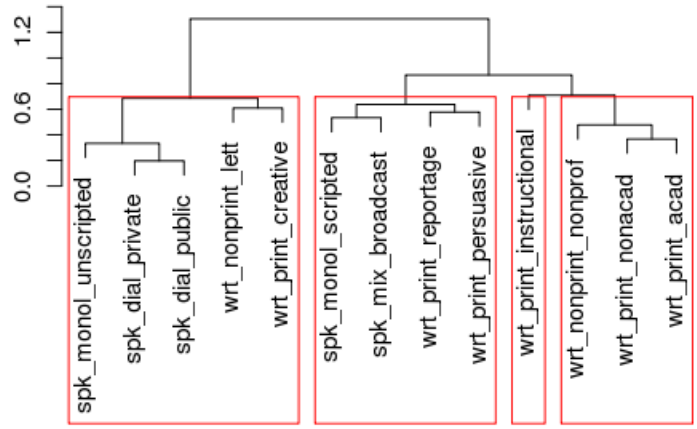

Figure 4: The most frequent 4-cluster solution for the 13 sub-registers of the ICE-GB

Interestingly, it is particularly the 4-grams that yield the least useful cluster solutions with seven clusters (the first 22 %-slices) or 10 clusters (all other larger slices), but while these have not much classificatory power, they are still not just noise as they preserve, for instance, the [spk] vs. [wrt] distinction.

3.6 The ICE-GB: interim summary

Again, there were several fairly clear trends, both in terms of content and methodology. With

regard to content there is again a robust and omnipresent distinction between spoken and written registers, even though the *n*-grams also indicate that some written data are very similar to spoken data, and it is exactly those kinds of findings that point to the need for a more fine-grained, multidimensional approach proposed by Biber and colleagues (if the resources for such labor-intensive analyses are available, that is). Spoken dialog is recognized better than spoken monolog, and on the whole, written printed registers prefer to cluster with other printed registers over non-printed ones. Importantly, even where the clusters as such are too coarse-grained, their internal structure is already indicative of the finer granularity larger *n*-grams as larger %-slices later reveal.

With regard to the methodology, again the cluster solutions at the MASWs are reasonable and *n*-grams show considerable discriminatory power: the cluster solutions, while not all completely homogeneous, fit well and correspond largely with the corpus compilers' decisions. In this corpus, the 1-grams are more compatible with the rest of the data, but the 4-grams are not as useful.

4. Concluding remarks

This paper explored four different questions, and we are now in a position to provide first answers:

i. How well can *n*-grams distinguish parts in corpora? Our answer is that *n*-grams alone can already discern substantial amounts of structure in the data, but that the *n*-gram-based approach here yields better results when applied to sub-registers than to registers, which may in itself be an interesting result since it might motivate corpus researchers in general to shift their focus more on the finer divisions of corpora than the fewer, more convenient coarser divisions.

ii. How much do corpus parts based on *n*-grams correspond to registers or sub-registers as defined by corpus compilers? All in all, our data yield results that are somewhat interpretable (for the coarse granularity of 4/5 corpus parts), but extremely well interpretable (for the finer granularity of 19/13 registers).

iii. Which *n*-gram length yields the best discriminatory power? In the context of the BNC-Baby and ICE-GB, it appears that 3-grams fare the best: 1-grams were generally volatile, 2-grams and 4-grams were occasionally off, and while neither 3-grams nor 5-grams exhibited erratic results, the former are computationally easier to handle and, thus, preferable (all other things being equal).

iv. How many of the most frequent *n*-grams yield the best discriminatory power? With the above caveat in mind, the data show that the relevant distinctions between the corpus parts emerge very early: with very few exceptions, a very small percentage of 3-grams resulted in dendrograms that correspond to the register distinctions made by human corpus compilers.

While these results are indeed highly compatible with, for instance, earlier studies by Biber and colleagues – by providing strong evidence in favor of the distinction between spoken and written data – they are still interesting because we obtained them:

− with clustering approaches that are more often used in the fields of information retrieval and/or word sense disambiguation, and not in linguistics;

- with using much fewer *n*-grams compared to most studies (given the small percentage slices yielding MASWs) and with a empirical approach to *n*-gram length (as opposed to an *a priori* defined *n*-gram length.);
- without any sophisticated grammatical analyses (à la Biber);
- without any keyword analysis based on a reference corpus (à la Xiao & McEnery).

While these initial results are very encouraging, there are aspects of this study that point r towards possible next steps. First, our results come from relatively well-studied but small corpora. Does the approach of saving computational effort by using only small slices of top 1% of *n*-grams result in similarly good results with larger corpora?

Second, our approach works much better when more sub-registers are provided as part of the gold standard, the input. On the one hand this may seem obvious (better input producing better output) but, on the other hand, it is not because, if a corpus is divided into more parts – sub-registers instead of registers – that also means that the type and token frequencies of *n*-grams will decrease considerably and that there are more ways for non-sensible clusters can arise. It seems, however, that the higher specialization of the sub-registers can offset the higher type/token frequencies of the broader registers.

Third, our results also show that sometimes even tiny changes in the data can result in very different numbers of clusters (yet with very similar structures), which should be taken as a warning: *n*-gram-based methods are very sensitive in that (i) they gloss over dispersion (cf. Gries 2008), (ii) they do not take corpus homogeneity or heterogeneity beyond the chosen level of granularity into consideration, and relatedly, (iii) just because corpus parts form some cluster structure based on *n*-gram frequency does not mean that a different phenomenon would not result in a very different cluster structure (cf. Gries 2006). Against this backdrop, researchers are advised to either take the above potential sources of noise into consideration or, minimally, do what we have done here, namely do not settle for one or two cluster solutions but rather use computational tools to explore the range of variation so that there is less of a chance that one or two outlier dendrograms distort the larger picture.

Be that as it may, in terms of the findings discussed here, we have shown that a pure *n*-gram-based approach can be used as an initial, computationally cheap, way of classifying corpus register that produces useful results. In terms of the methods, we have shown that the automated quality assessment of cluster solutions using average silhouette width is a useful heuristic to come to grips with the notoriously difficult problem of deciding on the number of clusters in large amounts of noisy data. We hope that our work stimulates more investigations of bottom-up register analysis and more varied exploration of methods for doing so.

References

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas. 1995. *Dimensions of register variation: a cross-linguistic comparison*. Cambridge: Cambridge University Press.

Biber, Douglas, Stig Johansson, Geoffrey Leech, & Edward Finnegan. 1999. *Longman Grammar of spoken and written English*. Harlow, England: Pearson Education Limited.

Biber, Douglas, Susan Conrad, & Viviana Cortes. 2004. If you look at …: lexical bundles in university teaching and textbooks. *Applied Linguistics* 25(3). 371-405.

Biber, Douglas, Eniko Csomay, James K. Jones, & Casey Keck. 2004. A corpus linguistic

investigation of vocabulary-based discourse units in university registers. In Ulla Connor & Thomas A. Upton (eds.), *Applied corpus linguistics: a multidimensional perspective*, 53-72. Amsterdam: Rodopi.

Cavnar, William B. & John M. Trenkle. 1994. N-gram-based text categorization. *Proceedings of SDAIR-94*, 161-75.

Chujo, Kiyomi. 2004. Measuring vocabulary levels of English textbooks and tests using a BNC lemmatised high frequency word list. In Junsaku Nakamura, Nagayuki Inoue, & Tomoji Tabata (eds.). *English corpora under Japanese eyes*, 231-249. Amsterdam: Rodopi.

Crossley, Scott A. & Max Louwerse. 2007. Multi-dimensional register classification using bigrams. *International Journal of Corpus Linguistics* 12(4). 453-78.

Csomay, Eniko & Viviana Cortes. 2010. Lexical bundle distribution in university classroom talk. In Stefan Th. Gries, Stefanie Wulff, & Mark Davies (eds.), *Corpus linguistic applications: current studies, new directions*, 153-168. Amsterdam: Rodopi.

Gries, Stefan Th. 2006. Exploring variability within and between corpora: some methodological considerations. *Corpora* 1(2). 109-151.

Gries, Stefan Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13(4). 403-437.

Hennig, Christian. 2009. Fixed point clusters, clusterwise regression and discriminant plots. Package for R 1.9 and higher. URL <http://www.homepages.ucl.ac.uk/~ucakche/>.

Kilgarriff, Adam. 2001. Comparing corpora. *International Journal of Corpus Linguistics* 6(1). 1-37.

Lee, David. 2001. Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle . *Language Learning and Technology* 5(3). 37-72 .

Mota, Cristina. 2010. Journalistic corpus similarity over time. In Stefan Th. Gries, Stefanie Wulff, & Mark Davies (eds.). *Corpus linguistic applications: current studies, new directions*, 67-83. Amsterdam: Rodopi.

Nishina, Yasunori. 2007. A corpus-driven approach to genre analysis: the reinvestigation of academic, newspaper and literary texts. *Empirical Language Research* 1. 1-36.

Orasan, Constantin & Ramesh Krishnamurthy. 2002. A corpus-based investigation of junk emails. Proceedings of LREC 2002, 1773-1780.

R Development Core Team. 2009. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing: Vienna. URL <http://www.R-project.org>.

Rousseeuw, Peter J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20(1). 53-65.

Santini Marina. 2007. Automatic Identification of Genre in Web Pages. Unpublished Ph.D. thesis University of Brighton.

Xiao, Zhonghua & Anthony McEnery. 2005. Two approaches to genre analysis: three genres in modern American English. *Journal of English Linguistics* 33(1). 62-82.

* The authors would like to thank Philip Dilts for input/discussion during earlier stages of the work culminating in this paper and an anonymous reviewer for helpful comments.

1 We use *register* here, following Biber (1995:9f.), as a situationally/communicatively-defined category of texts. However, our choice of *register* over one of the many somewhat similar and competing terms that corpus compilers (of the BNC, the ICE-GB, and others) as well as researchers have chosen (such as *genre*, *domain*, *text category*, and undoubtedly others), is essentially and intentionally theory-neutral; cf. Lee (2001) for a careful attempt to systematize this inventory of notions.

2 All retrieval and data management operations as well as all computations were performed with R (cf. R Development Core Team 2009).

3 Silhouette widths are computed on the basis of the ratio between the average dissimilarity of a clustered element *e* to all other elements in its cluster and the minimal average dissimilarity of *e* to other clusters; we embedded the function `cluster.stats` from the R library `fpc` (Version 1.2-4, Hennig 2009) into a script written by, and available from, the first author. Thus, silhouette widths fall between -1 and 1: well clustered elements have large positive silhouette widths whereas poorly-clustered elements have large negative silhouette widths, and a large average silhouette width for a cluster solution that groups *x* elements into *n* clusters indicates that the *n* clusters exhibit much within-cluster similarity and little between-cluster similarity (cf. Rousseeuw 1987).

4 Theoretically, it is of course possible that *n*-grams are very good at distinguishing registers but that the classification of the corpus files into registers by the corpus compilers was faulty. However, given the care that corpus compilers put into the selection of what to include in a corpus, we did not consider this an even remotely likely option and treated the corpus compilers' classification as the gold standard for our benchmark.

A reviewer pointed out that (what we call) registers are represented "poor[ly]" and differently in the corpora because "their constituent samples are either text fragments which either not contain all stages of the genre to which the text is said to belong, or hybrid texts which are highly likely to contain duplicate instances of genre stages." However, while we agree that, as in all sampling processes, one's sample(s) may represent the intended population to varying degrees, this does not affect the logic of the present approach. Our paper is concerned with how well our samples (proportions of differently long *n*-grams) recover corpus compilers' samples of the populations of registers, and as will be shown below, many of our *n*-gram samples yield very good results at recovering corpus compilers' samples even if those are not perfect approximations of the 'real' registers out there. Indeed, the very fact that our *n*-gram-based approach does as well as it does in spite of the different sampling processes used for the different corpora only underscores its robustness.