Bigrams in registers, domains, and varieties:
a bigram gravity approach to the homogeneity of corpora

Stefan Th. Gries
Department of Linguistics
University of California, Santa Barbara
stgries@linguistics.ucsb.edu

Abstract

In this paper, I explore how a new measure of collocational attraction proposed by Daudaravičius, & Marcinkevičienė (2004), lexical gravity *G*, can distinguish different registers and, hence, degrees of within-corpus homogeneity. To that end, I compute *G*-values for all bigrams in the BNC Baby and perform three different tests. First, I explore to what degree their averages reflect what is known about spoken vs. written data. Second, I use hierarchical agglomerative cluster analysis to determine how well a cluster analysis of the *G*-values can re-create the register structure of the BNC Baby (where I am using a classification of the BNC Baby into four registers and 19 sub-registers as a gold-standard). Finally, I compare the performance of the *G*-values in the cluster analysis to that of a more established measure of collocational strength, the *t*-score. The results show that the measure of lexical gravity not only distinguishes speaking and writing very reliably, but also reflects the well-known use of more frequent high-attraction bigrams in speech. Moreover, the gravity-based cluster analysis of the 19 sub-registers of the BNC Baby recognizes the corpus' register structure perfectly and, thus, outperforms the better-known *t*-score.

1. Introduction

For a variety of reasons, the corpus linguist's life is a very hard one because we have to constantly grapple with extreme variability. On the one hand, this is because the subject of interest is extremely variable: language and linguistic behavior are among the most variable phenomena studied by scientists because they are influenced by a multitude of factors which influence language only probabilistically rather than deterministically and which can be categorized into different categories:

general aspects of cognition having to with attention span, working memory, general intelligence, etc.;
specific aspects of the linguistic system: form, meaning, communicative pressures, etc.;
other performance factors (e.g., blood alcohol level, visual distractions, etc.).

On the other hand, the data on the basis of which we try to describe, explain, and predict linguistic behavior is very variable. While this is already true for (often carefully-gathered) experimental data, the situation of the corpus linguist using (often opportunistically-gathered) observational data is even more difficult for two reasons. First, corpora are only very crude samples of the real subject of interest, language, since they are

never infinite although language is in principle an infinite system;

never really representative in the sense that they really contain all parts or registers or genres or varieties of human language;

never really balanced in the sense that they contain these parts or registers or genres or varieties in exactly the proportions these parts make up in the language as a whole;

never complete in the sense that they never contain all the contextual information that humans utilize in, say, conversation; etc.

Second, in addition to all these imperfections, corpora are also extremely variable, not only in the ways in which they distort our view of the subject in the above ways, but also in how variability affects the quantitative data we obtain from corpora:

the variability of a frequency / percentage / mean: the larger it is, the less meaningful the frequency / percentage / mean;

the variability within a corpus (i.e., the within-homogeneity of a corpus): the larger it is, the more results from one part of the corpus generalize to the whole corpus or the (variety of the) language as whole;

the variability between corpora (i.e., the between-homogeneity of a corpus): the larger it is, the more the results from one corpus generalize to other corpora or the (variety of the) language as a whole.

These sources of variability are probably among the reasons why people after corpus-linguistic talks often ask “wouldn’t that be very different if you looked at another register/genre?” or “wouldn’t that be very different if you looked at another corpus?”, and of course the answer is always “sure, but the real question is, is it different enough to warrant this (other) distinction?”

Against this background, it is amazing and disturbing that there is relatively little work that systematically explores between- and within-corpus homogeneity. There are many studies that distinguish speaking and writing or a selected set of registers/genres but these studies usually already take these distinctions for granted rather than determining whether other divisions of the corpus would actually explain more of the variability within the corpus. Apart from Biber’s early word on corpus compilation (1990, 1993) as well as the large body of work by Biber and colleagues on the multidimensional approach to register variation (e.g., Biber 1988, 1995), there is little systematic exploration of these fundamental characteristics (which are related to the equally underexplored notion of dispersion) – some exceptions more or less concerned with these questions are Xiao & McEnery (2005), Santini (2007), Nishina (2007), Mota (to appear), Teich & Fankhauser (to appear) and references quoted therein – and there is even less work that explores these matters in a bottom-up fashion – some exceptions are Kilgarriff (2001), Gries (2005, 2006), Crossley & Louwerse (2007), and Gries et al. (2009).

One reason for this disturbing lack of studies is that any attempt to address this complex issue requires several tricky interrelated decisions:

on what level of granularity should the homogeneity of a corpus be measured? on the basis of the mode (e.g., spoken vs. written)? on the basis of registers (e.g., spoken dialog vs. spoken monolog vs. written printed, etc.)? on the basis of sub-registers (e.g., spoken private dialog vs. spoken public dialog vs. spoken scripted monolog vs. spoken unscripted monolog, etc.)? on the basis of corpus files (e.g., S1A-001)? …

which linguistic feature(s) is used to determine the similarity between the parts at some level of

granularity? characters? words? collocations? constructions? colligations? *n*-grams? …
how should similarity of some feature between parts at some level of granularity be computed/compared? raw frequencies or percentages? chi-square or log-likelihood? …

In this paper, I will try to take up some of these issues and address the homogeneity of a corpus, but I will differ from previous work in some ways

with regard to the level of granularity, I will explore the homogeneity of the corpus on more than one level to determine whether the differences between different corpus parts are in fact as substantial as is often assumed and/or which level of granularity is most discriminatory;
with regard to the linguistic feature studied: I will not just use raw frequencies or percentages or key words (which even requires a reference corpus), but bigram attraction;
with regard to how similarity is operationalized: I will use, and hence attempt to validate, a fairly new measure of collocational attraction whose merits have hardly been explored let alone recognized in the corpus community.

This approach and the validation of the collocational measure will be done in a bottom-up / data-driven way. The expectation is that, if the corpus is not completely homogeneous in terms of the registers that are supposedly represented in it and if the collocational measure works, then the measure should return meaningful register structure, meaningful in the sense that this structure should correlate with the corpus compilers' register decisions (unless their notion of register is ill-conceived). (Cf. Gries et al. (2009) for a similar exploration of how registers cluster depending on differently long and differently many *n*-grams.)

This paper has therefore two goals. First, it attempts to increase awareness of the fact that corpora can be divided into parts on many different levels, which is important because any decision to consider different corpus parts will have implications on the homogeneity of the results, or the lack thereof, so we must increase our understanding of this notion and its empirical consequences.

Second and more importantly, the paper attempts to undertake one of the first validations of a new measure of collocational attraction which, as I will show below, has a very interesting feature that should catapult it to the top of the to-look-at list of everyone interested in collocations.

The remainder of this paper is structured as follows. In the next section, I will discuss various aspects of the methodology employed here: which corpus was studied, how it was divided into different sub-registers, how the bigrams were extracted, which measure of collocational attraction was chosen and why, and how the corpus-internal structure was explored. Section 3 will discuss a variety of results following from these methodological decisions and will very briefly also compare the data based on the new collocational measure to an established measure, the *t*-score. Section 4 will summarize and conclude.

2. Methodology

In this study, I explore the collocational attractions of bigrams in registers and sub-registers of the British National Corpus Baby (<http://www.natcorp.ox.ac.uk>). This corpus exhibits considerable internal structure, which is represented in Table 1 and which is here based on the corpus compilers' decisions and also David Lee's more fine-grained classification. For the purposes of this study, this sub-division of the corpus into different parts is considered the gold-standard that any bottom-up exploration of the corpus should strive to recognize.

Mode	Register	Sub-register
spoken	demographic	AB, C1, C2, DE
written	academic	applied science, arts, belief/thought, natural science, social science, world affairs
	fiction	imaginative
	news	applied science, arts, belief/thought, commercial, leisure, natural science, social science, world affairs

Table 1: The structure of the BNC Baby

As mentioned above, this bottom-up exploration of the corpus will here be done on the basis of bigrams, which were generated as follows.[1] Each file was loaded, all non-sentences were stripped and all characters between word tags were extracted (regular expression: "<w[^>]*?>([^<]*?)</w>"). The data were then cleaned before further processing by removing many special characters (various kinds of brackets, other punctuation marks, asterisks, etc.) and numbers. Then, two sets of output files were generated, one with output files for each of the four registers and one with 19 output files (one for each sub-register). For within each (sub-register), these files contained the individual words, all sentence-internal bigrams, the number of the sentence in which each bigram occurred, and the complete sentences.

The measure of collocational attraction that I set out to explore and validate in this study is Daudaravičius & Marcinkevičienė's (2004) measure of lexical gravity *G*. For reasons not clear to me, this measure has so far not been validated or and hardly used although it has one very attractive feature that sets it apart from all other measures of collocational strength I am aware of (cf. Wiechmann 2008 for a comprehensive overview). Measures of collocational strength are generally based on 2×2 co-occurrence tables of the type represented in Table 2.

	word *y*	not word *y*	Totals
word *x*	*a*	*b*	*a*+*b*
not word *x*	*c*	*d*	*c*+*d*
Totals	*a*+*c*	*b*+*d*	*a*+*b*+*c*+*d*

Table 2: Schematic lexical co-occurrence table

The cell frequency *a* represents the same thing in all measures, namely the frequency of co-occurrence of *x* and *y*. However, for all measures other than *G* the frequencies *b* and *c* are the token frequencies of *x* where *y* is not and *y* where *x* is not respectively, but this means that the type frequency of cells *b* and *c* is not figured into the measure(s). That is, if *b*=900, i.e., there are 900 occurrences of *x* but not *y*, then all regular measures use the number 900 for the subsequent computation regardless of whether these 900 tokens consist of 900 different types or of 2 different types. The gravity measure, by contrast, takes this type frequency into consideration, as is indicated in (1).

(1) Gravity G ($word_1$, $word_2$) = $$\log\left(\frac{freq(word_1, word_2)\cdot type\ freq\ after\ word_1}{freq\ word_1}\right) + \log\left(\frac{freq(word_1, word_2)\cdot type\ freq\ before\ word_2}{freq\ word_2}\right)$$

This formula shows that, all other things being equal,

if freq($word_1$, $word_2$) increases, so does G;
if $freq\ word_1$ increases, G decreases;
if $freq\ word_2$ increases, G decreases;
if $type\ freq\ after\ word_1$ increases, so does G;
if $type\ freq\ before\ word_2$ increases, so does G.

This integration of type frequencies appears particularly attractive since it is well-known that type frequencies are in fact very important in a variety of areas:

(constructional) acquisition in first language acquisition (cf. Goldberg 2006: Ch. 5);
as determinants of language variation and change (cf. Hopper & Traugott 2003);
as a correlate of measures of (morphological) productivity (cf. Baayen 2001).

I therefore wrote scripts that computed lexical gravity values for all bigrams in the registers / sub-registers, a task which is computationally more intensive since one cannot just infer the frequencies in *b* and *c* on the basis of an overall frequency list but must look up the individual type frequencies in both slots for each of tens of thousands of bigram types in each of four registers and each of 19 sub-registers. However, in order to be able to compare the lexical gravity values with a much more established collocational measure, I also computed *t*-scores for each bigram according to the formula in (2).

(2) $$t = {observed\ bigram\ frequency}\Big/{\sqrt{\exp ected\ bigram\ frequency}}$$

Finally, once all gravity values for all bigrams in each register or sub-register were computed, they were analyzed in several ways that are summarized in Table 3. That is, the upper left cell means I computed the averages of the gravity values of all bigrams for each of the four registers. The upper right cell means I did the same but only for all bigrams that occurred more than 10 times. Then, I computed the average gravity of each sentence in each of the four registers. The same was then done for the 19 sub-registers. Finally, I clustered the 19 sub-registers based on the gravity values of each bigram type so that sub-registers in which bigrams are similarly strong attracted to each other would be considered similar. This cluster analysis based on the gravity values in the 19 sub-registers was then compared to a cluster analysis based on the *t*-scores. For both cluster analyses, I used the Pearson measure shown in (3) as the measure of similarity and Ward's method as the amalgamation rule.

Granularity	all bigrams	bigrams with n>10
4 registers	average *G* of each bigram type in each register	average *G* of each bigram type in each register
	average *G* for each sentence in each register	
19 sub-registers	average *G* of each bigram type in each sub-register	average *G* of each bigram type in each sub-register
	average *G* for each sentence in each sub-register	
		cluster analysis of the 4 registers based on the average *G* of each bigram type comparison to *t*-scores

Table 3: Kinds of exploration of the gravity values per (sub-)register of the BNC Baby

(3) $$\sum_{i=1}^{n}\left(freq_{part1}freq_{part_2}\right)\div\sqrt{\sum_{i=1}^{n}freq_{part_1}{}^{2}\cdot\sum_{i=1}^{n}freq_{part2}{}^{2}}$$

3. Results

In this section, I will report the results of the above-mentioned analyses For the analysis of the average tendencies, I will use box-whisker plots, for the cluster analyses I will of course use dendrograms.

3.1 The BNC Baby and its four broad registers

As for the average gravity values per register, the registers differ significantly from each other, which is little surprising on the basis of the sample sizes alone. It is more interesting to note that the main result is that the spoken data exhibit a smaller average gravity (both in terms of the median and the mean *G*-values) than all written registers. More specifically, the spoken register exhibits a mean gravity smaller than the overall mean whereas all written registers exhibit a mean gravity larger than the overall mean. This is true for all bigrams (cf. the upper panel in Figure 1) and for only the bigrams with a frequency of occurrence > 10 (cf. the lower panel of Figure 1).

This result for bigram *types* becomes more interesting when it is compared to the average gravity of bigram *tokens* per sentence in the same four registers, which is shown in Figure 2. The averages of the tokens per sentences show the reverse pattern: the average for the spoken data is highest, fiction is somewhat lower, and academic writing and news have the lowest values. Why is that? This is so because of several well-known characteristics of spoken language. While the four registers all consist of approximately 1m words, the number of sentences is much larger in speaking than in the three written registers. At the same time, we have seen above that there are fewer different bigram types in the spoken data. Thus, the tendency to have shorter sentences with more formulaic expressions (that have higher gravity values) leads to the high per-sentence gravity in the spoken data. The low values for academic writing and journalese, on the other hand, reflect the longer sentences that consist of more and less strongly attracted bigrams typical of the more elaborate and diverse writing.

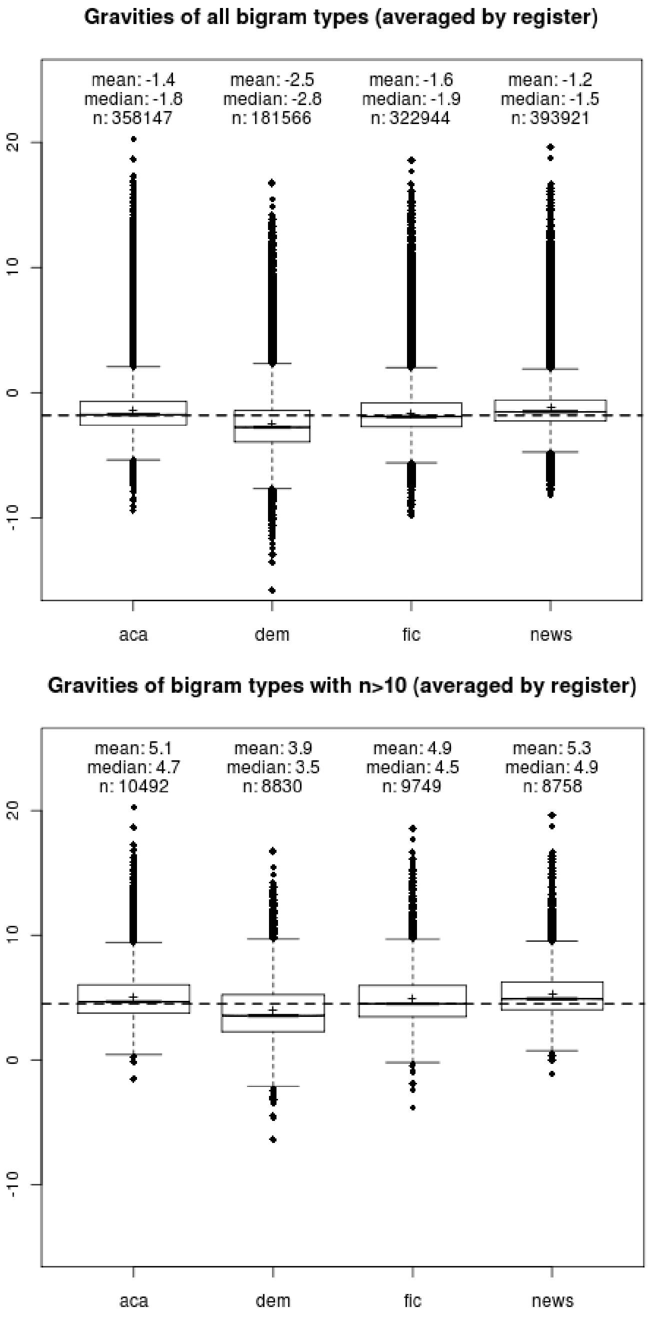

Figure 1: Box plot of average *G*-values (of bigram types) per register
(upper panel: all 2 grams; lower panel: frequent bigrams)

By way of an interim summary, the first results based on the *G*-values are reassuring and provide *prima facie* evidence for this measure. On the other hand, four registers do not exactly allow for a lot of variety in the results and stronger evidence from more diverse data would strengthen the case for gravity, which is why we now turn to the more fine-grained resolution of 19 sub-registers.

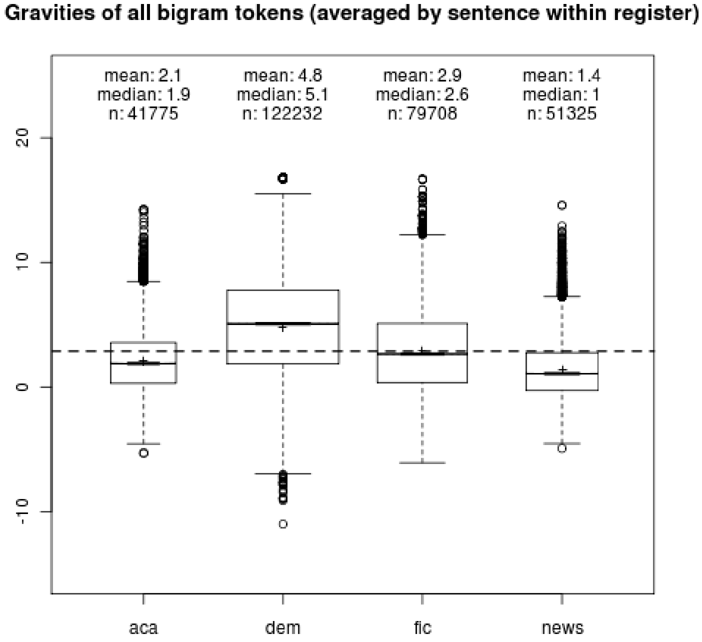

Figure 2: Box plot of average *G*-values per sentence per register

3.2 The BNC Baby and its 19 sub-registers

The results for the 19 sub-registers provide surprisingly strong support for the measure of lexical gravity *G*. For the sake of brevity, Figure 3 provides two kinds of results. The upper panel shows what the average *G*-values of all bigram types per sub-registers (i.e., what was the upper panel of Figure 1 for the four registers), but since the results for the frequent bigrams is for all intents and purposes the same, I do not show that here. Instead, the lower panel of Figure 3 shows the average *G*-values per sentence per sub-register (i.e., what was Figure 2 for the four registers).

Even at the much more fine-grained resolution of sub-registers, there is again a very clear and near perfect distinction of speaking vs. writing: again, the spoken data are characterized by low average gravities across bigram types and high average gravities per sentence. The only written sub-register that, in the lower panel, intrudes into an otherwise perfect spoken cluster is that one written (sub-)register one would expect there most: imaginative fiction, which can contain a lot of conversation in novels etc. and is often less complex in terms of syntactic structures etc.[2]

Within the written sub-registers, there is also considerable structure: For instance, in the lower panel the sub-registers of academic writing and journalese are separated nearly perfectly, too. Figuratively speaking, one would only have to move two academic sub-registers – belief/thought and arts – three positions to the right and would arrive at the gravity values perfectly recognizing that the 19 sub-registers are actually four registers.

Interestingly enough, the next analytical step reveals just that. The hierarchical cluster analysis of the 19 sub-registers (based on the frequent bigrams) results in a perfect register recognition; cf. Figure 4. The 19 sub-registers fall into two clusters, one containing all and only all spoken sub-registers, the other containing all and only all written sub-registers. The latter contains three clusters, which reflect exactly the three written registers distinguished by the corpus compilers. In addition, while imaginative fiction is clearly within the written cluster, it is also less written than academic writing and journalese. Also, even substructures within the academic-writing and the journalese clusters make sense:

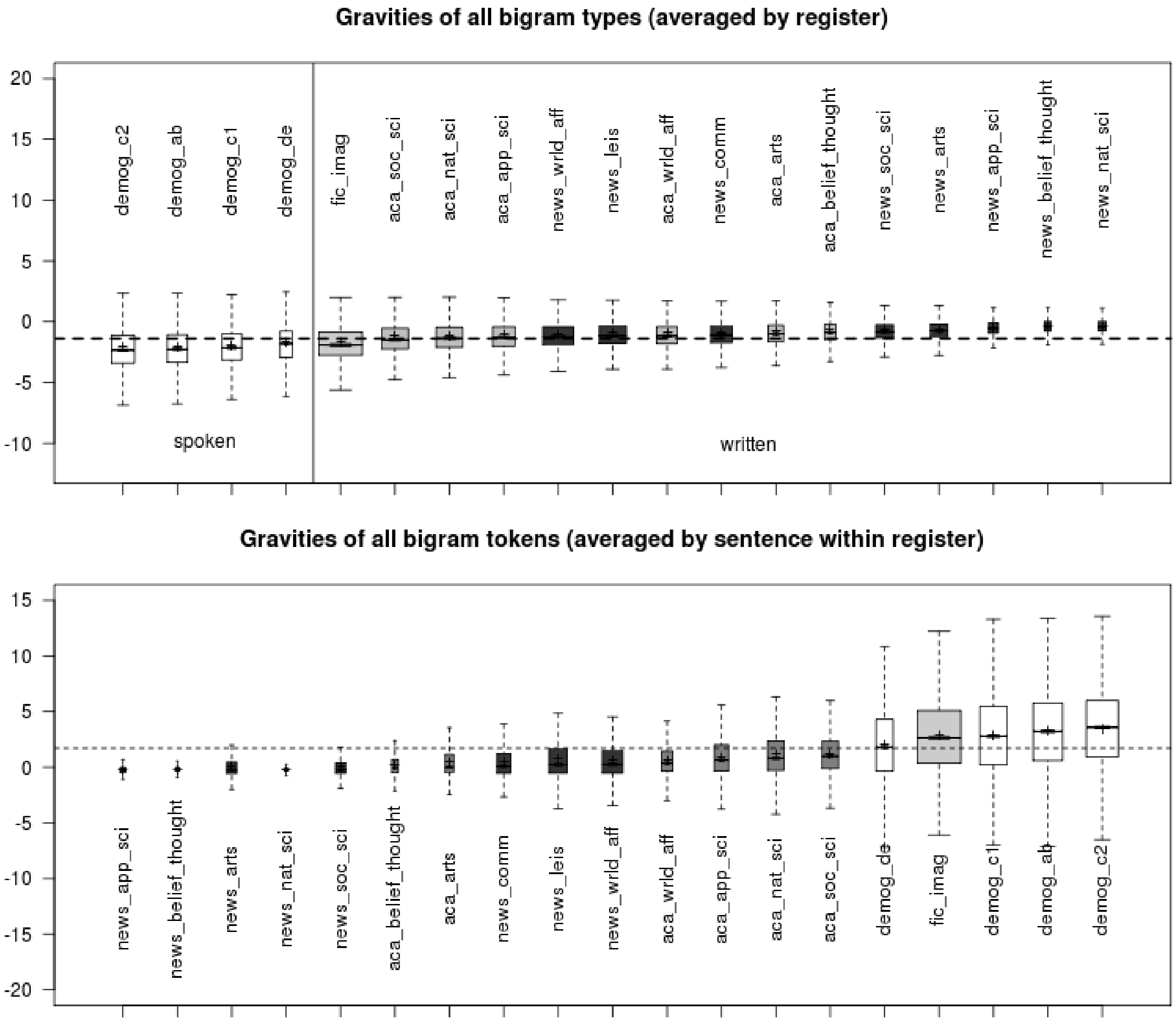

Figure 3: Box plot of average *G*-values (of all bigram types) per sub-register (upper panel)
Box plot of average *G*-values per sentence per register (lower panel)

in academic writing, arts and belief/thought are grouped together (as the more humanistic disciplines), then those group together with increasingly social-sciency data, then those group together with the natural/applied sciences: a nice cline from soft to hard sciences;
in journalese, the three sciences cluster together, as do arts and belief/thought.

It seems as if the gravity values are very good at picking up patterns in the data, given that the cluster analysis based on them returns such an exceptionally clear result. However, it may of course be the case that any collocational measure could do the same, which is why the gravity-based cluster analysis must be compared to at least one other cluster analysis. Consider therefore Figure 5 for the result of a cluster analysis based on the *t*-scores.

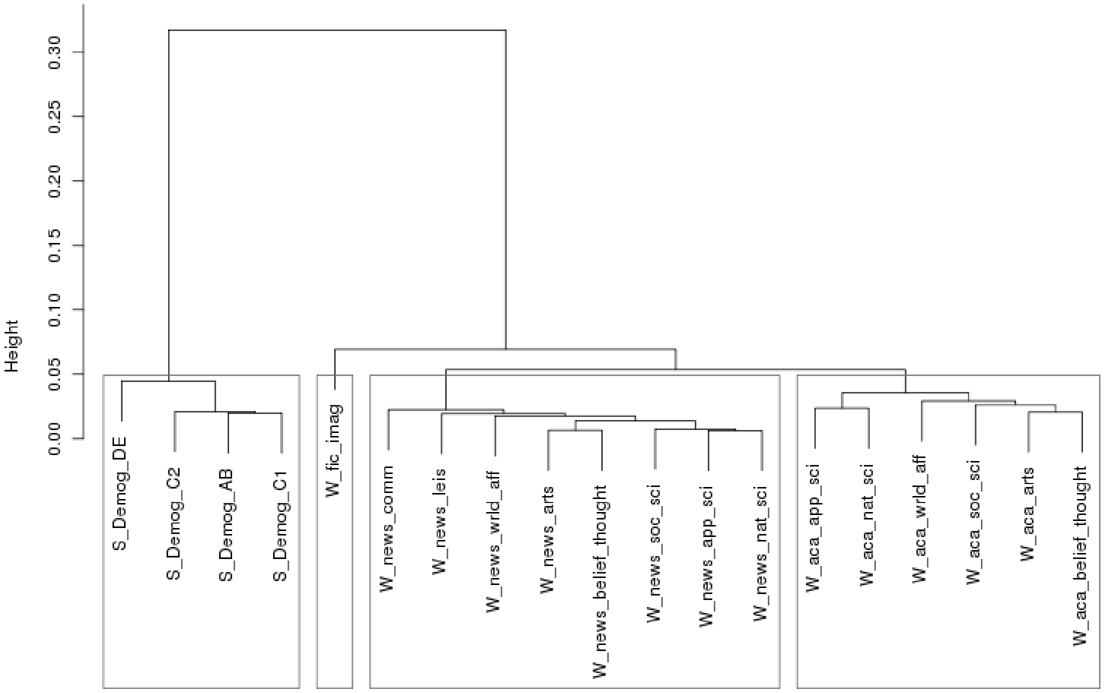

Figure 4: Dendrogram of the 19 sub-registers (based on the gravity values of all bigrams with a frequency larger than 10))

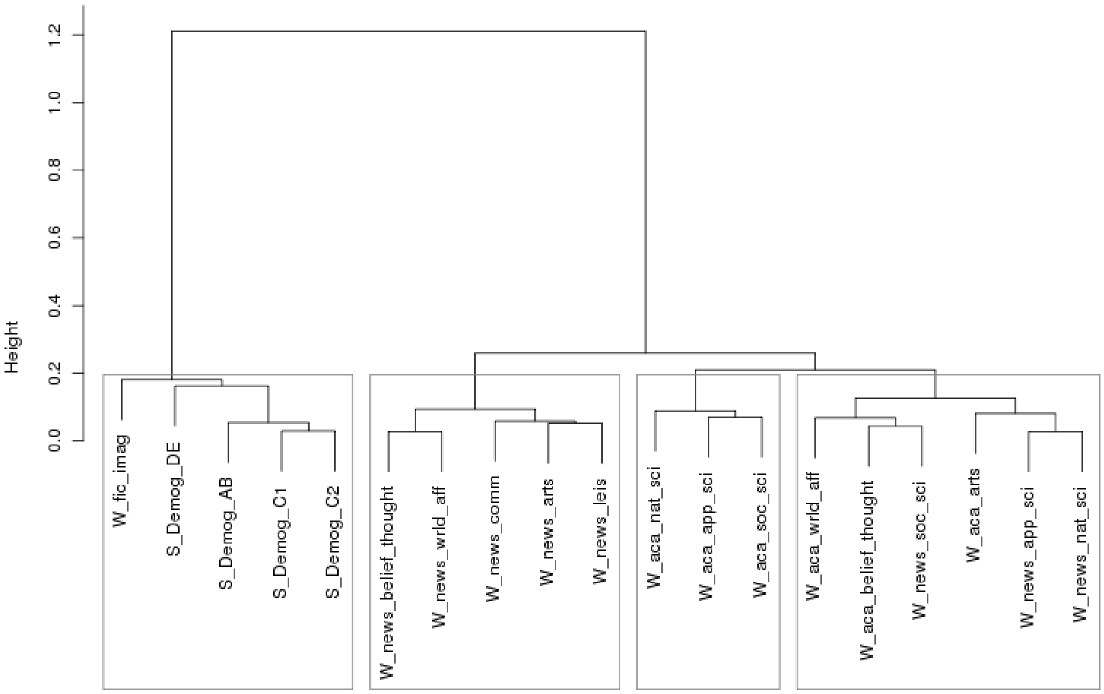

Figure 5: Dendrogram of the 19 sub-registers (based on the *t*-scores of all bigrams with a frequency larger than 10))

Obviously, this is also a rather good recognition of both the spoken vs. writing distinction as well as the four broad registers. However, it is just as obvious that this solution is still considerably worse than that of the *G*-values. First, spoken vs. written is not recognized perfectly because imaginative writing is grouped together with the spoken data. Second, there is one cluster that contains only journalese sub-registers, but not all of them. There is also a structure that contains all academic-writing sub-registers, but (i) this structure needs two separate clusters to include all academic-writing sub-registers (one of them at least contains all and only all sciences), and (ii) this structure then also contains three different journalese sub-registers. On the one hand, this is not all bad since, interestingly, it is the sciency journalese data that are conflated with the academic-writing sub-registers. On the other hand, two of the harder sciences are grouped together with a very soft-sciency academic sub-register. Thus, while the *t*-score dendrogram is certainly a good solution and even some of its imperfections are interesting and can be motivated *post hoc*, it is clear that the gravity-based dendrogram is much better at recognizing the corpus compilers' sampling scheme.

4. Concluding remarks

This paper pursued two different objectives, which we are now in a position to evaluate. In general, the results are nearly better than could have been hoped for. With regard to the issue of within-corpus homogeneity, there is good news for the compilers of the BNC Baby:

with the gravity approach, the register distinctions are strongly supported up to tiny sub-clusters within modes within registers within sub-registers
there is even some support from *t*-scores, but less clearly so.

Thus, the corpus compilers' assumptions of which registers to assume and which files to consider as representing a particular register are strongly supported. Put differently, the corpus exhibits exactly that internal structure that the register classification would lead one to expect. (This is of course not to say that a bottom-up exploration of this corpus on the basis of criteria other than bigram attraction could not lead to a very different result. As Gries (2006) stated, the homogeneity of a corpus can only be assessed on a phenomenon-specific basis.)

With regard to the issue of collocational measures, there is even better news for the developers of the gravity approach:

with the gravity approach, the register distinctions are strongly supported up to tiny sub-clusters within modes within registers within sub-registers (same point as above);
the cluster solution based on *G*-values clearly outperforms one very widely-used standard measure, the *t*-score;
the central tendencies of bigram tokens' gravity values per sentence match exactly what is commonly thought about speech: it uses highly cohesive chunks more frequently.

The high quality of the bigram-based cluster analysis is particularly interesting when compared to Crossley & Louwerse (2007:475) conclusion that

A bigram approach to register classification has limitations. While this analysis works well

> at distinguishing disparate registers, it does not seem to discriminate between similar registers […] Finally, while an approach based on shared bigrams seems successful, it is not an ultimate solution for register classification, but rather should be used in conjunction with other computational methods such as part of speech tagging, syntactic parsing, para-linguistic information, and multi-modal behavior […].

While I would not go so far as to say that a gravity-based bigram analysis is the "ultimate solution", the present results show clearly how powerful a solution it is and how well even sub-registers are clustered together. The present results, therefore, at least suggest that Crossley & Louwerse's call for the much more complex computational tools may be premature.

These findings have some implications not to be underestimated: Most importantly, the corpus-linguistic approach to collocational statistics should maybe be reconsidered, to move away from the nearly 30 only measures that only include token frequencies to one that also includes type frequencies. The type frequency-based measure of lexical gravity outperformed the *t*-score and, as mentioned above, it is well known that type frequencies are generally important in a variety of linguistic domains, which renders it somewhat surprising actually that it is only now that we are considering the possibility that type frequencies may also be relevant for collocations. This also means that, while the results reported here support lexical gravity, this does not mean that this measure cannot be improved any further. For example, the formula for *G* does not take the distribution of the type frequencies into consideration. If the type frequency of words after some word *x* is 2, then it may, or may not, be useful to be able to take into consideration somehow whether the two types after *x* are about equally frequent or whether one of the two types accounts for 98% of the tokens.

Another interesting idea is to extend gravities to *n*-gram studies. Daudaravičius & Marcinkevičienė (2004:333-334) propose to extract "statistical collocational chains" from corpora, successive bigrams with $G \geq 5.5$. In that spirit, Mukherjee & Gries (2009) used gravities to study how Asian Englishes differ in terms of *n*-grams: they

> computed *G*-values for all bigrams in their corpora;
> extracted chains of bigrams (i.e. *n*-grams) where all $G>5.5$;
> computed mean *G* for each *n*-gram; and crucially
> tested for each n-gram whether there is another *n*-gram that is one word longer and has a higher mean *G* – if there was no such longer *n*-gram, the shorter *n*-gram was kept, otherwise the longer *n*-gram was kept.

This approach is similar to Kita et al.'s (1994) approach to use a cost criterion as a bottom-up way to find differently long relevant *n*-grams and, maybe, opens up ways to identify different-length *n*-grams that are less computationally intensive than competing approaches involving suffix arrays etc. Given the initial promising results of the gravity measures and the important role this may have for our understanding and measurement of collocations, I hope that this paper stimulates more bottom-up genre analysis and more varied exploration of collocational statistics involving type frequencies and their distributions.

References

Baayen, R.H. (2001). *Word Frequency Distributions*. Dordrecht, Boston, London: Kluwer.

Biber, D. (1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. (1990). "Methodological issues regarding corpus-based analyses of linguistic variation". *Literary and Linguistic Computing*, 5, 257–69.

Biber, D. (1993). "Representativeness in corpus design". *Literary and Linguistic Computing*, 8, 243–57.

Biber, D. (1995). *Dimensions of Register Variation: A Cross-linguistic Comparison*. Cambridge: Cambridge University Press.

Crossley, S.A. and M. Louwerse. (2007). "Multi-dimensional register classification using bigrams". *International Journal of Corpus Linguistics*, 12, 453–78.

Daudaravičius, V. and R. Marcinkevičienė. (2004). "Gravity counts for the boundaries of collocations". *International Journal of Corpus Linguistics*, 9, 321–48.

Goldberg, A.E. (2006). *Constructions at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press.

Gries, St. Th. (2005). "Null-hypothesis significance testing of word frequencies: a follow-up on Kilgarriff". *Corpus Linguistics and Linguistic Theory*, 1, 277–94.

Gries, St. Th. (2006). "Exploring variability within and between corpora: some methodological considerations". *Corpora*, 1, 109–51.

Gries, St. Th., J. Newman, C. Shaoul, and P. Dilts. (2009). "*N*-grams and the clustering of genres". (Paper presented at the 31st Annual Meeting of the Deutsche Gesellschaft für Sprachwissenschaft).

Hopper, P.J. and E.C. Traugott. (2003). *Grammaticalization*. Cambridge: Cambridge University Press.

Kilgarriff, A. (2001). "Comparing corpora". *International Journal of Corpus Linguistics*, 6, 1–37.

Kita, K., Y. Kato, T. Omoto, and Y. Yano. (1994). "A comparative study of automatic extraction of collocations from corpora: mutual information vs. cost criteria". *Journal of Natural Language Processing*, 1, 21–33.

Mota, C. (to appear). "Journalistic corpus similarity over time". In St. Th. Gries, S. Wulff, and M,. Davies (eds.). *Corpus linguistic applications: current studies, new directions*. Amsterdam: Rodopi.

Mukherjee, J. and St. Th. Gries. (2009). "Lexical gravity across varieties of English: an ICE-based study of speech and writing in Asian Englishes". Paper presented at ICAME 2009, Lancaster University.

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

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>.

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

Teddiman, L. (2009). "Conversion and the lexicon: comparing evidence from corpora and experimentation". Paper presented at Corpus Linguistics 2009, University of Liverpool.

Teich, E. and P. Fankhauser. (to appear). "Exploring a corpus of scientific texts using data mining". In St. Th. Gries, S. Wulff, and M. Davies (eds.). *Corpus Linguistic Applications: Current Studies, New Directions*. Amsterdam: Rodopi.

Xiao, Z. and A. McEnery. (2005). "Two approaches to genre analysis: three genres in modern American English". *Journal of English Linguistics*, 33, 62–82.

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

[2] Fictional writing regularly takes an intermediate position between spoken data and journalese and/or academic writing; cf. Teddiman (2009) for the most recent example I am aware of.