

Benchmarks for text analysis: A response to Budge and Pennings

Kenneth Benoit^{a,*}, Michael Laver^b

^a *Department of Political Science, Trinity College, Dublin 2, Ireland*

^b *New York University, New York, USA*

1. Introduction

Budge and Pennings (2007) criticize the “Wordscores” method for computerized content analysis on essentially two grounds. The first is that the best test of Wordscores accuracy is whether it can “reproduce the rich time series produced by the MRG/CMP covering a 50 year period” (Budge and Pennings, 2007: 5), which Budge and Pennings claim it does not do. The second is that Wordscores time series estimates, as implemented by Budge and Pennings, yield very little variation around mean scores for the entire time series. In this brief response we make three simple points:

1. There is a fundamental and unresolved methodological problem with establishing the MRG/CMP time series as the “gold standard” against which to evaluate the accuracy of other estimates of party policy positions: namely, that there is no agreed method of assessing the uncertainty of MRG/CMP estimates. Yet not only is every number estimated in the MRG/CMP dataset generated by a single human coder—who are acknowledged to disagree with other real and potential coders,

introducing measurement error—but also the manifesto texts themselves represent stochastically generated verbal deposits of party positions whose random character is not represented in MRG/CMP scores. The net result of not having estimates of these forms of error associated with the MRG/CMP estimates means that it makes a fatally flawed benchmark, since it is impossible to know whether some independent estimate is the “same as” or “different from” the equivalent CMP estimate.

2. Another fundamental problem in this claim is that the CMP series is based on a coding scheme devised in the early 1980s, and this benchmark is probably moving over time. An equivalent manual-coding scheme devised in 1945, or 2005, would almost certainly be substantively different in significant ways, generating different results. In fact, this problem is no different from the fixed reference point required by Wordscores, the only difference being that the consequences of violating the assumption clearly emerge using Wordscores, yet are hidden by the CMP approach.
3. The Wordscores technique is misused by Budge and Pennings, in particular in their setup of reference texts and reference scores. In short, concatenating reference texts over a long period, is inappropriate for the task at hand and directly generates the “flattened” results they find. Averaged

* Corresponding author. Tel.: +353 1 608 2941; fax: +353 1 677 0546.

E-mail addresses: kbenoit@tcd.ie (K. Benoit), m1127@nyu.edu (M. Laver).

reference text scores, in other words, generate averaged virgin text score estimates—exactly the sort of “garbage in, garbage out” admonition underscored by Laver et al. (2003: 330).

2. The problem of unknown error in the CMP time series

Budge and Pennings make very strong claims for the CMP time series data, claims fruitfully examined in the light of the informative description of CMP coder reliability by Volkens (2007). Budge and Pennings claim that the CMP data “directly reflect what the parties state as their position rather than what others judge it to be” (Budge and Pennings, 2007: 10). Someone reading such a claim and coming to this debate for the first time might be surprised to discover that what the CMP data actually report, for a given manifesto, is what a single expert coder judged this manifesto to be saying, measured against the benchmark of the CMP’s 54-category coding scheme. “Manual” manifesto coding, *of its very essence*, reports what “others” (expert coders) judge party positions to be in the light of the words in the manifesto.

Users of the CMP dataset who have not carefully studied descriptions of how this was generated might be surprised to discover that every reported number in the dataset was generated by one human coder, once only. Each score for a manifesto on a coding category, therefore, comes with no estimate of associated error. This bears directly on the assertion that the CMP data are the gold standard with which all other types of estimate should be compared, since it is unclear how such a comparison can be rigorously effected, even if we assume CMP point estimates of policy positions to be completely unbiased. If the standard errors of the CMP estimates are very large, then almost any number generated by some other technique is consistent with these. If they are very small, then we have a greater possibility of discriminating between the two estimates. But we have no idea how large or small are the standard errors around the CMP estimates. Thus it is unclear how they can be used in a rigorous way to benchmark other measures.

This problem is greatly exacerbated when the virtues of a “rich time series” (Budge and Pennings, 2007: 5) are claimed for the CMP data. Since we have no estimate of the error associated with any CMP point estimate, it is quite unclear what to make of a time series of CMP point estimates. When two CMP estimates of the same party position differ between time points we have no way of knowing how much of that difference

can be attributed to the random error that must surely exist, and how much to a “real” underlying difference in policy positions. The comparison of CMP with Wordscores results from the Budge–Pennings paper illustrates the problem well: CMP is judged as performing better because it varies more, but we still have no way to know how much of this variation is real and how much is due to estimation variability and how much to fundamental uncertainty. Wordscores, on the other hand, combines a measure of estimation and fundamental variability in its standard errors, based on variances in word frequencies as well as the total observed number of words. (Budge and Pennings do not report these in their time-series comparisons, although this ultimately does not matter because of the more serious flaws characterizing this application of Wordscores, detailed below).

It is worth underscoring at this point the fundamental importance of having reliable measures of estimation uncertainty when measuring and reporting social and political phenomena. It is widely agreed that point estimates should always be accompanied by estimates of uncertainty, typically in the form of standard errors. So we view it as impossible that a quantitative measure for which no measure of uncertainty exists can be regarded as a “gold standard” for any area of empirical inquiry.

So what is the scale of the error in the CMP time series data? There is no way of knowing precisely, but orders of magnitude can be assessed using two useful pieces of information. Volkens (2007) discusses inter-coder reliability and notes judiciously that “there is no way of getting 100 per cent identical results with conventional content analytic approaches” (Volkens, 2007: 10). She reports results from a series of tests in which trained coders were assessed in terms of their ability to replicate a “master” coding of a single manifesto by the project’s designers. The “fit” of coders to the master coding is assessed very liberally, in terms of the *correlations* between coder and master in the percentages allocated sentences to the 54 coding categories. (Thus cancelling coding errors will not be assessed.) Discussing the performance of trained coders *who were receiving their second coding contract*, Volkens (2007: 10) reports a correlation of 0.85 between trained coders and the master coding. In other words, a trained CMP coder on a second contract on average retrieved only 72 percent of the information in the master coding, suggesting fairly high error in the CMP’s published estimates.

A second inkling about the level of error in the CMP data comes from the number of sentences coded for

Table 1
Lengths of texts analyzed by CMP

Total number of “quasi sentences” in text	Cumulative percentage of all CMP texts ($n = 1991$)
≤25	3.7
≤50	13.9
≤100	32.9
≤150	47.4
≤200	56.8
≤500	79.4

Source: CMP dataset, distributed on CD-ROM with Budge et al. (2001).

each manifesto, reported in Table 1 using data released by the CMP itself. Again, users of the CMP data may be surprised to see that 14 percent of all manifestos coded yielded less than 50 quasi-sentences, including uncoded sentences. For these manifestos, the number of coded sentences was less than the number of coding categories; some coding categories were thus constrained to score zero by virtue of the CMP’s own method. For fully one-third of the manifestos that form the basis of the CMP time series, the number of sentences was 100 or less. No sense is given in the reported CMP data of the extent to which the reliability and validity of the reported manifesto policy positions are impacted by the (sometimes very small) number of manifesto sentences available for coding. To produce the estimates ultimately reported and used by researchers, all CMP quasi-sentence frequencies are converted into proportions, but *without using any of the information about the quantity of quasi-sentence frequencies*. It is a bizarre and striking feature of the reported CMP estimates that, when there is more information (such as longer and more precise and comprehensive manifestos), this is not differentiated at all by their estimation procedure from when there is less information (such as short and non-informative manifestos). Wordscores, on the other hand, not only produces more precise estimates as manifestos increase in length and quality of content, but also represents this reduction in uncertainty through its standard errors.

There is a final source of error affecting policy estimates extracted from text analysis, stemming from the stochastic nature of the text generation process. The precise choice of words (or quasi-sentences) ultimately deposited in the form of election manifestos varies according to circumstance, author, resources, and perhaps national context, yet we assume that the policy positions that manifestos are used to measure are fixed. If we agree with this characterization—and we consider that no informed analyst would fail to—then the treatment of observed manifesto text as a non-random sample

puts quantitative text analysis at odds with all other applications in data-based statistical inference. And while Wordscores makes some use of this stochastic nature of text—by estimating uncertainty in part based on fundamental variances in word scores—it too ultimately skirts this issue. We thus view a promising, as yet unexplored, but ultimately necessary avenue for future research to be characterizing and representing the stochastic nature of text generation by actors whose policy preferences are fixed.

3. The problem of moving benchmarks in all policy time series

A large part of the Budge and Pennings critique of Wordscores is based upon a logical non-sequitur: “[t]he real promise of inductive computerized coding ... is its ability to process large amounts of text quickly and accurately. ... The most obvious way of checking whether it can is to reproduce the rich time series produced by the MRG/CMP”. It is of course a fallacy to infer that the ability to process a large amount of text implies the ability to reproduce a “rich” text-based time series. The primary virtue of Wordscores has always been its ability to process a huge amount of text generated by multiple authors—for example all speakers in a legislature. Wordscores is of its essence a cross sectional technique and the “test” constructed by Budge and Penning is logically spurious.

However, the reason why Wordscores should only be used with great care on documents from different time periods is in itself instructive with regard to *all* empirical time series data on policy positions. For Wordscores, the difficulty is that the political lexicon changes over time. Thus the same set of reference texts used to benchmark an analysis of policy positions at time t_i cannot validly be used to benchmark analyses at $t_i + j$ or $t_i - j$, unless the assumption is made that the political lexicon and its meaning is identical for the two time periods. (This problem is also at the heart of the methodological error made by Budge and Pennings in evaluating Wordscores; see below). This problem is self-evident for Wordscores, but a version of it also applies to all attempts to analyze text to generate long time series of policy positions, including that of the CMP.

The 54-category CMP coding scheme was designed in the early 1980s and reflected the best judgment of the political scientists involved about the high-dimensional policy spaces structuring party politics in the 19 countries analyzed at the time. The political meaning of this coding scheme has very likely changed over time, however, and it seems highly unlikely that even the

same set of political scientists, debating the issue in 1945, or 2005, would have produced the same coding scheme. In short the validity of the CMP coding scheme is unlikely to be time-invariant, rendering the validity of long time series generated using the CMP's scheme from the 1980s in this sense somewhat obscure.

This problem is greatly exacerbated for the CMP's left-right scale. Even if we accept that the validity of the CMP 1980s coding scheme is time invariant, the substantive coding categories making up the CMP's left-right scale was also a product of the 1980s. If the process of building this scale is carefully scrutinized, it can be seen to be an inductive product of a CMP data series that terminated in the 1980s. Thus coding categories included in the scale were those that loaded together in country-based exploratory factor analyses in the period 1945–1985.¹ In this sense the content of the CMP left-right scale is “centered” at 1965 or so. Changes in the political meaning of specific policy categories, both inside and outside the scale (with the environment and immigration leaping to mind) imply that the substantive “meaning” of the CMP (as with any other) left-right scale is likely changing over time. But it means that the very “strength” of the CMP dataset, which is the time series generated on the assumption that the categories are fixed over time, is also its principal weakness since we cannot really expect that these categories apply equally to all countries across all time points.

In fact, the overwhelming conclusion of independent work on measuring left and right policy positions is that the specific content of left and right is precisely *not fixed* across space and time. Inglehart and Huber's (1995) conclusion from a relatively open-ended expert survey to measure left and right in 45 countries was that the left-right dimension “is an amorphous vessel whose meaning varies in systematic ways” (Inglehart and Huber, 1995: 90) depending on national political, economic, and social context. Using the CMP coding categories, Gabel and Huber (2000) performed factor analysis on the CMP coding categories to extract a first principal component representing the left-right dimension, finding that its content varied considerably across countries and across time. The “most significant

lesson” from the study was “the importance of making as few *ex ante* assumptions as possible about the specific issues that constitute left-right ideology” (Gabel and Huber, 2000: 102)—precisely the opposite from the CMP approach to measuring left-right. Finally, Benoit and Laver (2007) showed from an extensive set of expert surveys conducted in 2002–2003 that the content of left and right also depends heavily on country context, and in many cases includes important issues such as the environment and immigration—two policy dimensions not included in the CMP left-right *ex ante* defined scale. The overall consequence of this problem of the left-right dimension as a moving target is that it renders irretrievably problematic the use of the CMP scale as the benchmark for assessing every other left-right scale in political science.

In fairness to the mammoth project that is the CMP, however, it is worth mentioning that the moving target problem is chiefly a shortcoming of the particular CMP left-right scale, not an intrinsic flaw in the coding category concept itself. Gabel and Huber (2000), for instance, show how the CMP codes can be used to construct an alternative left-right measure. Kim and Fording (1998) also explore alternatives to the CMP left-right scale that overcome other problems involved in its computation. Improvements on the CMP left-right scale are possible while still relying on the basic CMP data, in other words, although Budge and Pennings make no use of them.

4. Methodological problems in the Budge–Pennings implementation of the Wordscores technique

Up to this point we have taken issue with Budge and Pennings' claims that the CMP left-right scores serve as critical or even useful benchmarks for assessing the validity of other means of measuring left-right policy positions. Completely independent from our critique of the CMP measure, however, is the simple fact that Budge and Pennings have wrongly applied Wordscores in their time-series comparison of CMP and Wordscores left-right estimates. The essential problem lies in their concatenation of reference texts and in their averaging of CMP left-right scores to serve as reference values for Wordscores. In brief, the Wordscores procedure generates a list of words from chosen *reference texts*, based on the relative occurrence of each word across and within texts, given a set of reference scores assigned by the researcher for a given dimension—in this case, the left-right dimension. Point estimates on the original policy dimension are then generated for *virgin texts*,

¹ The left-right scale employed by the CMP was originally designed by Laver and Budge (1992) based on analysis of manifesto codings from the period 1945–1985. Their procedure collapsed the 54 sentence categories into 20 policy dimensions, established through a number of exploratory factor analyses to identify categories which consistently loaded together. Details on the CMP left-right scale are provided in Budge et al. (2001) and in Benoit and Laver (2007).

computed as the mean of the scores of the words in the virgin text, weighted by their relative frequencies within those texts. In addition to yielding point estimates for virgin texts, the procedure also computes confidence intervals.

One critical aspect of the Wordscores procedure is the choice of reference texts and the assignment of reference values. Reference texts must not only contain information on the policy dimension of concern, but also be lexically similar to virgin texts. Reference values, moreover, must discriminate adequately between a set of different reference texts. Reference texts must contain words that discriminate between different policy positions, since the results from the procedure rely on the selective association of certain words by text authors of certain policy orientations.

The error in the Budge–Pennings analysis is that it concatenates reference texts over a twenty-year period, and applies reference scores to those texts that are the mean positions of their dimension of interest, in this case the CMP left-right scores. According to Budge and Pennings, this aggregated document can “claim superior status to others” (Budge and Pennings, 2007: 10) as a calibrating instrument because they include all reference words likely to represent the political left-right content that Wordscores should be able to extract. But herein lies the problem. Any collection of words used as a reference text must be associated with a *discriminating* reference value, yet Budge and Pennings employ *averaged* CMP left-right scores as reference values. Each bundle of words for each reference text is then associated with this average value (or more precisely according to their analysis, the twenty-year average minus the virgin text being scored). When virgin documents are scored, then, it should come as no small surprise that the estimates for the virgin texts are also average values. Indeed, using such a procedure is *guaranteed* to produce flat times series, with the only difference between party estimates being associated with the average positions over the time period—not individual changes at different time periods.

An excellent illustration of the problem is in fact provided by Budge and Pennings themselves in footnote 9, where they report encountering the extreme version of this misuse of reference values by also having tried including the virgin texts in the concatenated reference texts. This procedure yields *completely* flattened estimates, since all of the words in each virgin text are associated perfectly with the averages for the entire twenty-year period! The only possible way that it can be otherwise is if words in a virgin text are also found in other reference texts, something which also happens,

but which in practice (with reference texts of this size) only shrinks the variation in the virgin estimates “almost to the vanishing point” rather than to zero.

Budge and Pennings are correct in that a key conclusion in the Wordscores procedure “is certainly that the a priori score dominates the process” (Budge and Pennings, 2007: 18). This is exactly what we warned in the initial presentation of the Wordscores method, that

the word scores for each policy dimension, and hence all subsequent estimates relating to virgin texts, are conditioned on the selection of reference texts and their *a priori* positions on key policy dimensions. This is thus something to which a considerable amount of careful and well-informed thought must be given before any analysis gets under way. In this, our method shares the “garbage in-garbage out” characteristic of any effective method of data analysis; potential users should, indeed, be comforted by this. (Laver et al., 2003: 330)

In other words, Budge and Pennings have demonstrated that Wordscores works exactly as advertised, by showing that estimates for policy texts associated with similar texts whose reference value is a twenty-year mean, produce results pretty close to the twenty-year mean. The fact that the mean values for each party (e.g. the British example) differ in the ways that we expect is in fact reassuring, since at least it proves that word frequencies alone can indeed differentiate parties on the basis of information they are supplied, namely their 20-year averages.

5. Conclusions

We have contested the use of the CMP left-right scale as a benchmark for evaluating other measures of left-right policy positions because it provides no associated estimates of uncertainty, making it impossible to tell whether differences between the CMP and other scales are real or due to error. Furthermore, we have argued that the pre-defined and fixed CMP left-right scale cannot truly measure the real political content of left and right over time and across countries, because this content varies considerably across time and space. Differences between the CMP measures and other estimates are likely due to the fact that for different countries and different time periods, therefore, they are simply measuring different things.

Independently from these general points about the validity of CMP, we have also demonstrated that the Budge–Pennings implementation of Wordscores to generate left-right estimates for comparing with CMP scores is fundamentally flawed. They put garbage in,

and get garbage out—a feature of any honest method of producing estimates of any kind. The Budge–Pennings analysis, we must conclude, does not represent any reasonable attempt to “validate” computerized word frequency estimates.

References

- Benoit, K., Laver, M., 2007. Estimating party policy positions: comparing expert surveys and hand-coded content analysis. *Electoral Studies* 26 (1), 90–107.
- Budge, I., Klingemann, H.D., Volkens, A., Bara, J., Tannenbaum, E., Fording, R., Hearl, D., Kim, H.M., McDonald, M., Mendes, S., 2001. *Mapping Policy Preferences: Parties, Electors and Governments: 1945–1998: Estimates for Parties, Electors and Governments 1945–1998*. Oxford University Press, Oxford.
- Budge, I., Pennings, P., 2007. Do they work? Validating computerised word frequency estimates against policy series. *Electoral Studies* 26 (1), 121–129.
- Gabel, M., Huber, J., 2000. Putting parties in their place: inferring party left-right ideological positions from party manifesto data. *American Journal of Political Science* 44, 94–103.
- Inglehart, R., Huber, J., 1995. Expert interpretations of party space and party locations. *Party Politics* 1, 73–112.
- Kim, Heemin, Fording, Richard C., 1998. Voter ideology in western democracies, 1946–1989. *European Journal of Political Research* 33 (1), 73–97.
- Laver, M., Benoit, K., Garry, J., 2003. Estimating the policy positions of political actors using words as data. *American Political Science Review* 97, 311–331.
- Laver, M., Budge, I., 1992. *Party Policy and Government Coalitions*. St. Martin’s Press, New York.
- Volkens, A., 2007. Strengths and weaknesses of approaches to measuring policy positions of parties. *Electoral Studies* 26 (1), 108–120.