Available online at www.sciencedirect.com
ScienceDirect
Electoral Studies 26 (2007) 121–129
Electoral Studies
www.elsevier.com/locate/electstud

Do they work? Validating computerised word frequency estimates against policy series

Ian Budge [a,1], Paul Pennings [b,*]

[a] University of Essex, Colchester, UK
[b] Department of Political Science, Vrije Universiteit Amsterdam, De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands

Abstract

Policy scorings of political actors are crucially important in operationalising rational choice models and other important theories in political science. Doing them more cheaply and quickly by computer is important for the advancement of the discipline. But we can hardly substitute these for hand-coding or even use them in new fields without being sure of their validity and reliability. We check this by comparing the mappings produced by word frequency methods with the policy series available from the work of the Manifesto Research Group/Comparative Manifesto Project (MRG/CMP). Using an aggregate calibrating/reference 'document set' for the time period in question evades reliability problems with pairwise comparisons and provides an authoritative text which enables individual party platforms to be scored and mapped over long time periods. Comparisons of the techniques for two countries (US and UK) are not encouraging. Wordscores in their current operationalisation flatten out party movement just as previous computerised approaches have done. Sensitivity testing with British party manifestos 1979–1997, using an expert scoring, does not reveal any improvement in performance. The reliability problems which arise with policy series are also likely to recur with cross-sectional applications of the word frequency approach.

Keywords: Political texts; Hand coding; Party manifestos; Word frequencies; Computerised content analysis; Time series; Reliability; Policy estimates; Comparative analysis

1. Background

Understandable enthusiasm is generated by the prospect of computerising political textual analyses, with enormous savings in the time and costs of manual coding, and a great extension in their political applications. Texts, written by participants themselves, are after all the prime way of establishing their policy positions and codifying policy outcomes. Other indicators of policy position represent indirect ways of getting at what texts state directly.

The 50 year policy series produced for parties, electorates and governments (Budge et al., 2001) are thus one of the research successes of recent years. Unfortunately, attempts to computerise the coding have not matched up, possibly because they impose too many restrictions on the words to be counted (Bara, 2001; Laver and Garry, 2000). Kleinnijenhuis and Pennings (2001) (KP) broke out of these with their pioneering

[*] Corresponding author. Tel.: +33 20 5986852; fax: +33 20 5986820.

E-mail addresses: budgi@essex.ac.uk (I. Budge), pjm.pennings@fsw.vu.nl (P. Pennings).

[1] Tel.: +44(0)1206 872499; fax: +44(0)1206 873234.

0261-3794/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.electstud.2006.04.002

proposal to associate (almost) all the words in a 'calibrating' set of documents with their known scorings, and use their frequencies of association with such scorings to estimate the scores of documents in an 'application' set.

The scores KP were interested in were the quasi-sentences already counted manually into each of the coding categories devised by the Manifesto Research Group (MRG) (Budge et al., 2001, 215–219). Their procedure (Budge et al., 2001, 164–168) was then as follows:

(i) A selected set of party programmes formed the 'calibrating set'. Each quasi-sentence in each programme had already been assigned by a human coder to one MRG/CMP category, giving an overall percentage distribution and a score for each category.
(ii) A set of fifty-six probabilities was then assigned to each word occurring in the calibrating set: one probability for each of the fifty-six categories of the MRG scheme. The extent to which a given word is an indicator of a given MRG category depends on the inductive probability of it occurring in the fifty-six MRG categories in the calibrating set.
(iii) Words that occurred less than five times were removed from the dictionary. Words that occurred extremely often, but did not discriminate between the categories, were removed also, for example, function words like 'the'.
(iv) The probabilistic dictionary derived in this way was applied to the 'application set' of party programmes to be coded. For a complete party programme, the 'frequency' of a specific MRG category in that party programme can be computed as the sum of the frequencies of the words from the calibrating set in the application set, weighted by their respective probabilities of pointing towards the MRG category under review.

The calibrating set in their analysis consisted of three of the five Dutch party programmes of 1998 (PvdA, VVD, CDA). The application set was all other Dutch party programmes in the period 1946–1998. The estimated percentages were then aggregated for each programme in the 'application set' to form scores on a Left-Right scale. An overall 'map' of Dutch party movement over time was formed from the estimated scores and compared with the map based on original scores (like Fig. 1). The two maps were broadly similar, locating parties in basically the same positions in regard to each other over time. However, enough party locations (and party moves from one point in time to another) did *not* correspond, for the authors to conclude that their word frequency approach did not work well.

Laver, Benoit and Garry (LBG) (Laver et al., 2003) adapted KP's approach to deal with a single overall score for each 'calibrating,' or in their terminology 'reference,' document. Dealing with a single score allowed them to estimate the probability of association of each word with the scores in terms of their frequency of occurrence in the reference documents. They operationalised their procedure in a computer programme Wordscore: http://www.tcd.ie/Poliical_Science/wordscores/. Surprisingly, as their approach was also a word-frequency one, LBG ignored KP's negative evaluation of their research and went through a different set of tests. Using expert judgements rather than direct estimates of policy position from the MRG/CMP codings, they compared these with economic

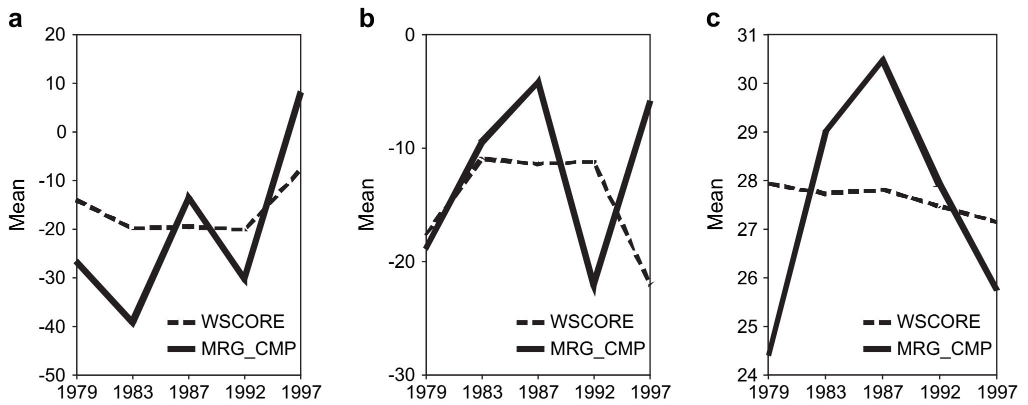

Fig. 1. Manifesto and Wordscore estimates of party positions and movements in UK 1979–1997. Solid line shows MRG/CMP, dashed line shows Wordscore estimates. Scales run from −100(left) to +100(right). The different figures have different metrics. With a unit metric Conservative Wordscore estimates are numerically close to Manifesto ones, whereas Labour ones are far away. (a) L-R estimates for the Labour Party; (b) L-R estimates for the Liberal Party; (c) L-R estimates for the Conservative Party.

and social scores estimated by Wordscore from 1992 British and Irish manifestos on 1997 ones. In three out of four of these comparisons the Wordscore estimates for 1997 matched those made by experts. The technique also produced plausible scoring results for Germany 1990–1994 (where later expert scorings were not available). An additional test applied the technique cross-sectionally, to characterise Irish legislators' degree of support for the government in a confidence debate. These results led LBG to evaluate the word frequency approach very positively.

The divergence of opinion between the main proponents of the word frequency approach needs to be resolved. One thing however is clear: if word frequency analysis is to replace manual coding in its most frequent field of application, it must be able to create valid time series of the sort the Comparative Manifesto Project has produced. These are clearly richer and more useful for analyses of party strategic behaviour (Adams, 2001; Budge, 1994) than scattered pairwise comparisons can be. As over time comparisons have been the main tests applied to the computerised scoring, they, in any case, provide the most obvious check on whether the word frequency approach produces valid results in general. We go into more detail about this in Section 3. Prior to that, however, we consider problems built into the procedure of pairwise comparison and how they can be tackled.

2. Problems of pairwise comparison

2.1. Building long time series

The real promise of inductive computerised coding, as LBG rightly point out, is its ability to process large amounts of text quickly and, hopefully, accurately. The most obvious way of checking whether it can do this is to reproduce the rich time series produced by the MRG/CMP, covering a 50 year period for many democracies. KP realise this, which is why their test tries to reproduce the whole map of post-war Dutch party movement. Comparisons limited to two time points in the 1990s seem inadequate to check an approach which has ambitions for mass processing of texts.

Quite apart from the limited test they provide, pairwise comparisons also seem an inadequate empirical base for generating large sets of scores. Suppose we start with reference texts and a priori scores for 1997. We then estimate scores for virgin manifestos for 2001. Do we then use these to estimate scores for virgin manifestos in 2005, and these for 2010, and so on? The potential for cumulative error here seems very large.

Even if we conceive of replicating time series for the past, the same problem occurs if we think of it in terms of pairwise comparison. Of course, we could always start off each pair with a known independent estimate. But if we already have a series of independent estimates available there is little point in having a computerised analysis. The point is to *replace* other estimates. But if we are to do this on a large scale, pairwise comparisons seem inadequate.

Perhaps, however, we are only thinking of limited point estimates rather than extended time series. Even in terms of estimating one set of positions from another, however, problems arise in terms of the precedence and authority of the texts involved, which we now consider.

2.2. Does looking forward coincide with looking back?

President Bush's policy position could be characterised as slightly more to the Left in 2004, compared to 2000, or as having been slightly more to the Right in 2000, compared to 2004. We normally think of these statements as saying the same thing. This is because we make the implicit assumption that there is an invariant Left-Right scale on which the 2000 and 2004 positions can be located. Comparing positions forwards or backwards thus makes no difference.

In *estimating* these positions through word frequencies, particularly if these use an invariant scale, forwards and backwards does make a difference, however. Typically, expert surveys produce the same judgements about party positions, whatever the date at which they were collected. In their careful analysis of the major expert surveys to date, McDonald and Mendes (2001: 124–130) conclude that 'the experts are reporting an over-time general statement about party locations in the post-war period'. This is because they are essentially locating parties on the scale presented to them in terms of their membership of the traditional party families (McDonald et al., 2007).

The general nature of expert judgements (whatever the strictures in the questionnaire about judging where the parties are *now*) makes it legitimate and possible to apply them to any set of party documents, not simply those collected at the date nearest the expert survey.[2] This raises, in acute form, the compatibility of forwards and backwards estimates. The same

[2] Rather than a weakness, their generalisability could be regarded as a strength,of course, when such surveys are only available for a limited number of time points. Müller and Strøm (2000) use Laver and Hunt's (1992) placements to characterise parties back to 1950. See also Kitschelt (1994).

scores for Left-Right positions of the parties will be applied to 2000 or 2004 document sets with differing word frequencies. With each set taken in turn as the calibrating/reference one they are very likely to produce different estimates, since words will be assigned different probabilities of association with the same score positions in a backwards as compared to a forwards comparison.

At first sight, it may seem more natural to estimate forwards, as we are proceeding from a better known to a lesser known situation. LBG take it as so natural that they do not even discuss the assumption. Doing things in this sequence is less obvious than it seems, however, since KP, in the other major computer use of word frequencies, use 1998 Dutch manifestos as the basis for estimating scores for previous ones—an extended series of backwards-looking pairwise comparisons. In any case, we would feel doubtful about a 2000 on 2004 estimate if it subsequently turned out differently from a 2004 on 2000 estimate. Which is the more reliable? To avoid reliability problems we would need to have a guarantee built into the approach that backwards and forwards estimates would coincide. This point would still apply even if we had different sets of expert placements of parties for the two time points involved, as LBG have for Britain and Ireland in 1992 and 1997. (But then, we might ask again why we need computerised estimates if we have got other valid ones already.)

We deal with this problem in an over time framework here, like everybody else, because of the availability of independent scores. However, the question of the authority and precedence of the 'calibrating' or 'reference' documents also applies to cross-sectional analyses, as we discuss in our conclusions.

3. Pairwise comparisons: empirical checks

Table 1 checks out these possibilities with British data for the 1992 to 1997 comparisons. We cover both the situation in which individual document scores are available separately for 1992 *and* 1997 (MRG/CMP estimates for individual party manifestos) and the situation in which an invariant expert placement of the parties is made (Castles and Mair, 1984).

The original MRG/CMP scorings for 1992 and 1997 are: −30.40 and 8.07 for Labour; −22.10 and −5.86 for Liberal; and 27.90 and 25.74 for Conservative. There is a wide consensus among specialists and commentors on these moves: a dramatic Rightwards shift by 'New Labour' from a relatively Left wing position in 1992, a more modest Rightwards move by Liberals from 1992 to 1997, and little change by Conservatives from a strongly Rightist position.

Table 1
'Forwards' and 'backwards' pairwise comparisons of Left-Right movement of British Party manifestos 1992 and 1997[a]

	Labour	Liberals	Conservatives
(a) *Using CMP/MRG individual estimates for 1992 and 1997*			
'Forwards' comparison			
1992 on 1997			
1992 CMP/MRG reference score	−30.4	−22.1	27.9
1997 Wordscore estimate	−14.2 (1.78)	−34.85 (1.84)	27.1 (1.66)
'Backwards' comparison			
1997 on 1992			
1992 Wordscore estimate	1.89 (1.03)	−1.52 (0.92)	27.46 (0.71)
1997 CMP/MRG reference score	8.07	−5.86	25.74
(b) *Using Castles and Mair expert scoring*			
'Forwards' comparison			
1992 on 1997			
1992 expert reference score	2.30	5.00	7.80
1997 Wordscore estimate	4.12 (0.18)	2.98 (0.18)	8.21 (0.16)
1997 on 1992			
1992 Wordscore estimates	2.61 (0.27)	4.52 (0.21)	8.63 (0.16)
1997 expert reference score	2.30	5.00	7.80

[a] Bracketed entries are standard errors for the standardised Wordscores. 'Forward' comparisons in the table take the individual MRG/CMP scores for 1992, and the invariant Castles–Mair expert scores, associate these with the 1992 manifestos as reference documents, and estimate scores for the 'virgin' manifestos of 1997. 'Backward' comparisons take the individual MRG/CMP scores for documents in 1997, and the invariant Castles–Mair expert scores, associate them with the 1997 manifestos as reference documents, and estimate scores for the virgin documents of 1992. MRG/CMP scores run from −100 (Left) to +100 (Right). Castles–Mair expert scores run from 0 (Left) to 10 (Right).

What do the Wordscore estimates tell us about party changes between these years? Looking forwards from 1992, Labour is correctly characterised as moving Right in 1997 on the basis both of Manifesto and expert scores. Looking backwards from 1997 to 1992 on the basis of the MRG/CMP scores, Labour is also seen as being to the Left in 1992 compared to 1997—but on the basis of very different policy positions. Looking forward from 1992 on the basis of the expert scores, Labour is also estimated to have moved Right: from 1997 however, Labour is judged to have been more to the Right in 1992.

These discrepancies are the most interesting ones in Table 1 because Labour's sweep rightwards was the most dramatic feature of the 1997 election. However, we should note that the Liberals were also generally thought to have moved a bit right, whereas three estimates in Table 1 have them moving strongly Left. The Conservatives, generally considered to have moved little or veering very slightly Left, are shown as such on three estimates, but as moving Right on the expert scorings from 1992 on 1997.

As a whole, Table 1 shows considerable discrepancies between pairwise comparisons, depending on whether these are forwards or backwards looking. Sometimes they directly contradict each other. The contradictions are more marked with the invariant expert scoring, as one would expect, of course. But even starting from the individually scored MRG/CMP data one gets a very different impression of where Labour and Liberals are moving from and where they are going, depending on whether one looks forwards or backwards.

Such difficulties derive from the lack of an authoritative document set that is clearly the reference or calibrating one. Individual time-bound sets of documents in a series cannot usually claim superior status to others. However, one can be *created*, quite simply, by aggregating relevant documents for the time period under consideration (e.g. the late post-war). The word frequencies of the aggregated documents can then be associated with expert scorings (as they, in effect, relate to the whole period) or with mean Left-Right scores from the Manifesto Data for the relevant period, which can be used as input to Wordscore. We do both below.

This procedure also has the advantage that we can map the estimated scores for all parties against the MRG/CMP time series, thus providing the essential, but till now absent, check on the validity and performance of the computerised approach over different countries and time points.

4. Manifesto estimates and wordscores: bases for comparison

The suitability of the MRG/CMP time series for evaluating word frequency estimates is clear. They are, in fact, the only text-based time series we have, apart from the few point estimates of party positions that have been made using other techniques. Moreover, they directly reflect what the parties state as their position rather than what others judge it to be. Their validity and reliability have been extensively examined (Budge et al., 1987, 2001; Laver and Budge, 1992; Klingemann et al., 1994), and they have been used satisfactorily by a variety of authors (Adams, 2001; Baron, 1991, 1993; Blais et al., 1993; Schofield and Parks, 2000; Erikson et al., 2002; Warwick, 1994, 2002).[3]

In terms of their relevance to checking word frequency estimates, we have already noted that the pioneers in this area developed the approach as a computerised operationalisation of MRG coding procedures (Kleinnijenhuis and Pennings, 2001). The affinity between the basic logic of the Wordscore programme using word frequencies, and measurement procedures for the MRG Left-Right scale, is shown in Table 2. This is doubly relevant because both the scale and the programme are central to our empirical comparisons and evaluations below. Wordscore is the generator of the word frequency estimates with the Left-Right mean score as its a priori input, while the individual Left-Right scores from the MRG/CMP coding are used to evaluate the success of the computerised estimates.

In the Wordscore procedure, word probabilities have been adapted to estimate positions on an extraneously given scale, on which parties have already (again extraneously) been located at a given point in time. Of course, the extraneous scale can be the MRG Left-Right scale, and potentially any policy scale derived from the MRG/CMP data or from expert or other judgements.

One obvious resemblance, in light of the relative emphasis approach which dominates the MRG/CMP coding (Budge et al., 2001: 75–92), is the fact that Wordscore relies on the relative frequencies of words, ignoring completely their positive or negative

[3] An odd criticism has been derived from erroneous citations of test results for coders in training (the failures then being retrained) as final figures for trained coders (Laver et al., 2003: 317 (fn6 and fn17); Laver and Garry, 2000). Other checks are cited in Budge et al. (1987, passim). Figures for the stability of data structures are reported in Budge et al. (2001, 111–141).

Table 2
Similarity of the procedures used in the MRG/CMP Left-Right scorings and in computerised word frequency approaches to scoring manifestos[a]

Stages	MRG/CMP Procedures	Wordscore
1. Extraneously given scale	Theoretically driven groupings of coding categories to form the Left-Right scale	A priori scale assigns numeric score to each party document in calibrating/reference set
2. Weighting procedure	Each coding unit (category) in scale weighted equally in absence of extraneous theoretical advice on how to weight otherwise	Each coding unit (word) weighted in accordance with its frequency over the calibrating/reference set manifestos
3. Scoring of new (virgin) manifestos	Party scores calculated by adding for each manifesto the percentaged frequencies of quasi-sentences in each category in Left and Right Groupings and subtracting Left sum from Right sum	Party scores are calculated by adding weighted word scores for each manifesto in the application (virgin) set

[a] References: Budge et al. (2001: 21–22), Laver et al. (2003: 314–319).

connotations, in establishing estimates of party locations. As noted by the authors (Laver et al., 2003, 329–330):

> This … has to do with the way words are used in practice in the advocacy of particular policy positions. With regard to our own technique take the individual word used in our earlier example – “choice”. Of course, the word “choice” has several meanings, while each meaning can also be qualified with a negative or even a double negative. Someone coming to computational text analysis for the first time might reasonably feel for these reasons that the relative frequency of the word “choice” in a given text does not convey substantive information … our approach works because particular words do, empirically, tend to have policy-laden content. Thus, in post-Thatcher Britain those using the word “choice” in relation to education or health policy, for example, tended to be advocating greater choice of schools or health providers and correspondingly less central control. Those opposing such policies tended, as a matter of empirical observation, not to argue for “no choice” or “less choice but rather to talk about the benefits of central planning and coordination.

An explicit illustration of the underlying identity of Wordscore and CMP procedures is given in Table 2. This divides the procedures into three stages:

1. Selection of a scale that is ‘given,’ in the sense of existing before the procedures are applied. In the MRG/CMP case this is given from pre-existing ideological writing, which stresses certain themes as important to either Left or Right ideology. The full set of opposing themes, operationalised as MRG/CMP coding categories, then form the two ends of the scale. The Wordscore procedure simply takes a pre-existing set of scale scores that could be produced either by experts or by the MRG/CMP or by other means.
2. The coding unit to be used in the estimation is weighted. The weighting is a priori in the case of the MRG/CMP; ideological discussions do not indicate how each theme or category is to be weighted in relation to others, so, by default, they are given equal weights. Obviously, if a relevant theory did attribute a greater weight to some categories they could be weighted differentially. The computerised procedure estimates weightings for each word empirically, on the basis of its frequency in each of the calibrating/reference set of manifestos.
3. The weighted frequencies of the coding units in the application set (the ‘virgin’ documents) are added to form an estimated scale score for each party. In the case of Wordscore this can be compared with the original scale position to see how far the party has changed policy. The MRG scores can, of course, be calculated independently for each party manifesto and compared.

Procedures, therefore, differ in detail between two techniques, but share the same underlying logic in proceeding from extraneously given scale to weightings to calculation of final scores. This gives us the opportunity of comparing computerised estimates with the well-attested manual scores, with a view to validating the former.

5. Empirical comparisons of Wordscore estimates using the MRG/CMP Left-Right scoring as extraneous input and empirical check

Our procedure, therefore, is to input the Left-Right mean score, as estimated by MRG/CMP for the relevant time period, into Wordscore. The reference texts used to provide weightings for words are the aggregated manifestos for each party over the time period involved, omitting the individual manifesto whose scores are being estimated.[4] Wordscore is then used to compile estimates of Left-Right position for each individual party programme. These can be mapped onto a figure (cf. Fig. 1) for comparison with the original MRG/CMP scores for each document. The latter have been extensively validated against historical accounts (Budge et al., 2001, 19–50) and in other contexts (see above). Thus they form a good criterion of the validity of the computerised estimates.

We illustrate this procedure for British party manifestos 1979–1997. This forms a single political period marked off by strong Conservative dominance. The stability of the overall political situation makes it a good one for sensitivity testing of Wordscore, which we report below. There are, moreover, considerable contrasts between the parties in terms of their policy profiles over the period. The Conservatives stuck to a relatively extreme Right-wing position. The Liberals fluctuated a bit in the Centre Left. Labour fluctuated markedly from a fairly Leftist position in 1979 to 'the longest (Left-wing) suicide note in history' (1983), to a much more Centre-Left position in 1987, a move back Left in 1992 and a dramatic dash to the Centre in 1997. These tendencies are captured by the MRG/CMP mappings in Fig. 1.

What do the Wordscore estimates show? They certainly carry on from the a priori mean MRG/CMP estimates input to the programme in making a consistent differentiation between the parties. However, compared to the individual MRG scorings they drastically flatten out policy movements. The most obvious example of this is with Labour. The dramatic policy changes from 1979–1983–1987 are imperfectly reflected (in the case of 1983 and 1987 not reflected at all). Nor is the distinctly Leftist stance of 1992, with its emphasis on the social services, really contrasted with the more Rightwards stance of 1987. The dramatic move to centre right in 1997 is also underplayed.

The same could be said of Liberal policy movements, where two seem wrongly signed by the estimates. Wordscore does catch the constancy of the Conservatives to a strong Right-wing position. But this may be just a happy chance of the limited variation of estimates about the mean.

The US case illustrates the limitations of the Wordscore estimates based on mean positions even more strongly (Fig. 2). These hardly move around the mean, whereas the Manifesto estimates catch the dramatic repositioning of the Democrats under Clinton in 1992, as well as less notable moves.

These estimates were supplemented by others using the aggregated sets of reference documents with the Castles-Mair expert party placements as the a priori input. Not surprisingly, these flattened out the path of the parties even more than the Manifesto means, and are not reported here.

6. Overview

The word frequency approach to analyses of political texts is currently the most promising basis for their computerisation, offering a comparatively simple and direct way of estimating their political position. The extent to which this builds on the selective emphases approach dominant in manual coding of texts and responsible for building up the only long policy time series available has been underplayed. But it is important in providing a conceptual basis for comparisons between them.

Indeed, our major focus has been on re-orientating the word frequency approach from an exclusive concern with out of context point estimates (what happened in 1997 relative to 1992) to its original goal of creating policy time series comparable with those available from manual codings. Sooner rather than later computerised analyses will have to come to grips with this problem, since their usefulness for mass processing of texts

[4] The 'virgin' texts are omitted in each case from the aggregated 'reference' documents because there is a possibility that, if they were included, their word frequencies would not differ much from those in the 'reference documents' because they occur in both. In fact, when we did aggregate all documents for the time period into the 'reference set,' overall variation in the estimates was attenuated almost to vanishing point. To get the results reported below, therefore, we associate MRG/CMP mean left-right scores for the time period 1979–1997 inclusive in Britain, with the following aggregate reference sets for each 'virgin' set:

Virgin set	Aggregate reference set
1979	1983 + 1987 + 1992 + 1997
1983	1979 + 1987 + 1992 + 1997
1987	1979 + 1983 + 1992 + 1997
1992	1979 + 1983 + 1987 + 1997
1997	1979 + 1983 + 1987 + 1992

The same procedure was used for the US (Fig. 2). Although the set of documents used as the reference set changes with the year under investigation, it still represents general tendencies in manifestos over the period in question, and is clearly a more stable and authoritative base than any one year's set of manifestos would be for estimating policy positions in an individual election. In fact, the estimates still flatten out, as can be seen below, but this is more clearly an empirical result. We thank Ken Benoit and Mik Laver for making this point.

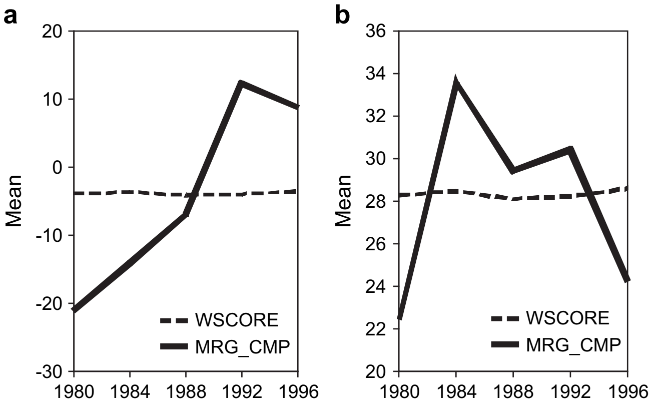

Fig. 2. Manifesto and Wordscore estimates of party positions and movements in US 1980–1996. Solid line shows MRG/CMP, dashed line shows Wordscore estimates. Scales run from −100(left) to +100(right). The different figures have different metrics. With a two-unit metric Republican Wordscore estimates are numerically close to Manifesto ones, whereas Democratic ones are far away. (a) L-R estimates for the Democratic Party; (b) L-R estimates for the Republican Party.

will be judged on whether they can match the richness and informativeness of manual coding in this regard.

Putting things in a long-term perspective immediately raises questions about the suitability of pairwise comparisons for creating time series. Indeed, extending our time perspective even to the next election raises questions about the authority and primacy of the two sets of texts involved and whether looking backwards produces the same results as looking forwards. Our analysis indicates that there is no guarantee that the two will coincide—a conclusion which then raises questions about the overall reliability of the pairwise procedure even in a cross-sectional context.

An alternative which solves the problem of the texts' authority and primacy, and thus of what a priori to use as a base for the word frequency analysis, is to use some score relating to the whole of the time period under review, with a set of texts which also have that quality. This is most easily provided by aggregating texts for the time period involved and using them as the calibrating reference set. Simply using a series of pairwise comparisons (1979–1983, 1983–1987, 1987–1992, 1992–1997) presupposes that there *is* a parallel series of estimates to feed in at each point, and thus eliminates the need for computerisation to replace it. It also begs the question of why look forwards rather than backwards, or how to resolve discrepancies if we do both.

Alas, when we adopt the alternative approach of aggregating texts to use with the Manifesto Left-Right mean, or expert placements of parties, as a priori input, we do not get much variation of individual manifestos around it. Manually based time series are more informative and accurate. This may suggest going back to the original concern of Kleinnijenhuis and Pennings (2001) with developing an automated coding system, with more internal differentiation than one gets by associating a single score with a whole document.

One methodological conclusion about the current incarnation of the word frequency approach in WORDSCORE is certainly that the a priori score dominates the process. Results rest very much on the initial differentiation of the documents it introduces, and may not get very far from this in the course of analysis (cf. Figs. 1 and 2). The small differences uncovered also raise the old question of whether statistically significant differences between document scores are necessarily very important ones substantively.

7. Cross-sectional analyses using word-frequencies

This paper has applied word-frequency approaches to over-time analyses. Not only does this cover a major potential application, it also follows the precedent set by both KP and LBG, five of whose six texts involve time. The latter have, however, intimated verbally that they see the main application as being to the processing of many texts cross-sectionally. Do the same reliability problems apply here? It seems that they do, whenever no document set has clear precedence over the others under analysis to form the reference texts.

Laver et al. (2003, 327) failed in attempts to set up manifestos as authoritative reference texts for Government policy documents,[5] which might have solved the

[5] For a bridging of this gap see Budge et al. (2001, 245–250).

problem. The actual cross-sectional analysis they do report, of an Irish confidence debate (Laver et al., 2003, 327—328), simplifies by having pro and anti-Government scores applied to speeches by the Prime Minister, the coalition partner's leader, and the leader of the main opposition party. In many countries, however, the party Chairman is an equally or more authoritative party figure than the leader pro-tem. Would estimates based on their pronouncements or on those of institutionalized factions within the party necessarily coincide with those based on statements of the official leadership? In a cross-sectional context, looking backwards rather than forwards finds its equivalent in looking from one side rather than another. Until there is some guarantee that differently-based estimates will coincide, we have a stability and reliability problem.

To sum up, the word frequency approach has great promise, but the jury is still out on whether it will be fulfilled. One crucial test will be its ability to create valid policy series to match those currently produced by manual coding.

References

Adams, J., 2001. A theory of spatial competition with biased voters. British Journal of Political Science 31, 21—23.

Bara, J., 2001. Using manifesto estimates to validate computerised analyses. In: Budge, I., Klingemann, H.D., Volkens, A., Bara, J., Tanenbaum, E. (Eds.), Mapping Policy Preferences: Estimates for Parties, Electors & Governments 1945—1998. Oxford University Press, Oxford, pp. 143—156.

Baron, D., 1991. A strategic bargaining theory of government formation in parliamentary systems. American Political Science Review 85, 137—164.

Baron, D., 1993. Government formation & endogenous parties. American Political Science Review 87, 34—47.

Blais, A., Blake, D., Dion, S., 1993. Do parties make a difference? Parties and the size of government. American Journal of Political Science 37, 40—62.

Budge, I., 1994. A new spatial theory of party competition: uncertainty, ideology and policy equilibria viewed comparatively & temporally. British Journal of Political Science 24, 443—467.

Budge, I., Robertson, D., Hearl, D., 1987. Ideology, Strategy and Party Change: Spatial Analyses of Post-War Election Programmes in 19 Democracies. Cambridge University Press, Cambridge.

Budge, I., Klingemann, H.D., Volkens, A., Bara, J., Tanenbaum, E., 2001. Mapping Policy Preferences: Estimates for Parties, Electors & Governments 1945—1998. Oxford University Press, Oxford.

Castles, F., Mair, P., 1984. Left right political scales: some expert judgements. European Journal of Political Research 12, 73—88.

Erikson, R., Stimson, J., McKuen, M., 2002. The Macro Polity. Columbia University Press, New York.

Kleinnijenhuis, J., Pennings, P., 2001. Measurement of party positions on the basis of party programmes. In: Laver, M. (Ed.), Estimating the Policy Positions of Political Actors. Routledge, London, pp. 101—182.

Klingemann, H.D., Hofferbert, R., Budge, I., 1994. Parties, Policies & Democracy. Westview, Boulder, CO.

Kitschelt, H., 1994. The Transformation of European Social Democracy. Cambridge University Press, Cambridge.

Laver, M., Hunt, W. Ben, 1992. Policy and party competition. Routledge, London.

Laver, M., Garry, J., 2000. Estimating policy positions from political texts. American Journal of Political Science 44, 619—634.

Laver, M., Benoit, K., Garry, J., 2003. Extracting policy positions from political texts using words as data. American Political Science Review 97, 311—330.

Laver, M., Budge, I. (Eds.), 1992. Party Policy & Government Coalitions. Macmillan, London.

McDonald, M., Mendes, S.M., 2001. Checking the party policy estimates: convergent validity. In: Budge, I., Klingemann, H.D., Volkens, A., Bara, J., Tanenbaum, E. (Eds.), Mapping Policy Preferences: Estimates for Parties, Electors & Governments 1945—1998. Oxford University Press, Oxford, pp. 127—142.

Müller, W., Strøm, K., 2000. Coalition Governments in Western Europe. Oxford University Press, Oxford.

Schofield, N., Parks, R., 2000. Nash equilibrium in a spatial model of coalition bargaining. Mathematical Social Science 39, 133—174.

Warwick, P., 1994. Government Survival in Parliamentary Democracies. Cambridge University Press, Cambridge.

Warwick, P., 2002. Toward a Common Dimensionality in West European Policy Spaces. Party Politics 8, 101—122.