Available online at www.sciencedirect.com
ScienceDirect
Electoral Studies 26 (2007) 136–141
Electoral Studies
www.elsevier.com/locate/electstud

Missing the message and shooting the messenger: Benoit and Laver's 'response'

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

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

Benoit and Laver (BL; 2007-a-b) ignore the main point made in our article (Budge and Pennings, 2007), about the unreliability of policy estimates made by word frequency methods in the absence of authoritative 'calibrating' texts. Instead they concentrate on the general unreliability of the MRG/CMP Manifesto data (Budge et al., 2001) as a 'benchmark', in spite of the fact that we also used expert scores (Castles and Mair, 1984). Accordingly we divide this reply into two parts. The first deals with questions BL avoid but which provide a context for their criticism of aggregated reference documents. The second responds directly to their data-critiques, showing that they are mostly partial and misleading.

1. The general instability of word frequency estimates

BL (p. 1) characterize our article as 'criticizing the 'Wordscores' method for computerized content analysis …'. Actually it does not. 'Wordscores' is a useful computer programme[1] which operationalizes a particular approach within the range of possible word frequency techniques. But it is not the only approach. Indeed 'Wordscores' was clearly conceived within the conceptual framework of calibrating and application documents and word weightings provided by Kleinnijenhuis and Pennings (1999, 2001) and described at the outset of our article. The Laver team abandoned an earlier and cruder approach to computerized analysis (Laver and Garry, 2000) in favour of simplifying and operationalizing this word frequency one.[2] So we have at least two extant applications of word frequency analysis to political texts.[3]

The fact that we began with, and returned to, Kleinnijenhuis and Pennings's work at various points demonstrates our general concern with word frequency analysis as an approach rather than Wordscores as such. A check against the Manifesto Left-Right series was Kleinnijenhuis and Pennings's preferred test for their own word-frequency approach. So it was natural

* Corresponding author. Fax: +33 20 5986820.
E-mail address: pjm.pennings@fsw.vu.nl (P. Pennings).

[1] Indeed as we made heavy use of it to support our points about the word frequency approach, we ought to commend the speed with which its authors have made it generally available and in particular Ken Benoit's role in supporting and disseminating it among the professional community.

[2] Laver et al. (2003, 312) list Kleinnijenhuis and Pennings's article among earlier and quite disparate computerized content analyses but do not refer to it otherwise, or in their response, in spite of its prominence in our discussion.

[3] Word frequency approaches antedate both sets of authors of course. For example, a project in the Edinburgh Faculty of Divinity in the mid-1950s tried to establish the exact authorship of St Paul's Epistles through word counts. It stopped after concluding they could not all have been written by the same person but could not decide where St Paul was author because they lacked an authoritative reference text indubitably by him. This is the general problem we identify for word-frequency policy estimates.

0261-3794/$ - see front matter
doi:10.1016/j.electstud.2006.03.010

to extend the restricted pairwise comparisons which Laver et al.'s simplified variant had undergone.[4]

Immediately one starts thinking about word frequency analyses in the context of policy time series however certain conceptual queries arise which would enable one at a pinch to dispense with data-based analyses entirely in favour of thought-experiments. For example, what is the basis - the calibrating or the reference documents - on which to characterize the 'application' or 'virgin' documents? For Kleinnijenhuis and Pennings (2001) the answer seemed simple - take the most recent set of party documents (1998) and use them to characterize all the other party manifestos 'backwards'. For Laver et al. the answer was equally obvious: use the earlier documents (1992) to characterize the later ones (mostly 1997) 'forwards'.

Of course, as becomes evident on comparing the two cases, the choices are equally arbitrary. There is no reason why either documents from 1992 or from 1998 should be more authoritative than any other set of documents in the comparisons.

This in turn prompts the query, if neither 'backwards' or 'forwards' comparisons can be authoritatively preferred, would they provide the same estimate? Only if that could be guaranteed would we avoid a conflict of policy position estimates every time they are made, and with them a massive reliability problem. We should stress once again that this is not created by Wordscores, an honest operationalization of a particular word-frequency approach. Rather it is generated by the logic of the word-frequency approach itself, as the parallel examples of Kleinnijenhuis and Pennings (2001) and St Paul's Epistles (fn 3) show.

What BL should worry about are not the graphs of party movement in our article (Figs. 1 and 2 of Budge and Pennings, 2007) but Table 1 (also in Budge and Pennings, 2007) on 'forwards' and 'backwards' estimates which conflict. Incidentally, these are more severe when we use Castles and Mair's (1984) expert scores than when we use the Manifesto scores. It seems that convergence cannot be guaranteed between the estimates made on different bases. The same problem, as we noted, applies to cross-sectional estimates. Looking from one, or both, sides 'in' is not guaranteed to produce the same estimate as looking from the centre 'out'. In this context the standard errors produced by Wordscores *once it is given a set of reference documents to work on* give no inkling of the uncertainty produced by the fact that these documents are not definitive. This is a 'pre-Wordscores' problem but it fatally infects the policy estimates it produces.

Seen in this light we were actually on Laver et al.'s (and Kleinnijenhuis and Pennings's) side when we tried to create an indisputably authoritative set of calibrating or reference documents by aggregating documents over a given period! BL in their Section 3 argue very strongly that this was mistaken. If it is a case of 'garbage in–garbage out' however, word frequency analyses are buried in it too unless other authoritative sets of reference documents can be found. Nor do BL provide (or even try to provide) an answer.

We ourselves however are not entirely convinced about aggregation necessarily flattening out all useful information in the calibrating texts. Means are after all useful in identifying outliers. If there is strong variation among the application documents they will surely differentiate themselves from the mean. Actually Fig. 1a,b in Budge and Pennings (2007), for the British Labour and Liberal parties, show Wordscores tracing out quite a lot of movement on their part. The Conservatives (in Fig. 1c of Budge and Pennings, 2007) are shown as fairly static but that is probably correct. Were we only conducting a pairwise comparison of 1992–1997 as LBG did (2003, 319–321) the lurch rightwards by Labour shown in Fig. 1a of Budge and Pennings (2007) would be taken as a success for the Wordscores comparison. So would the move leftwards from 1979 to 1983. There is no indication here that a general flattening in reference document variation buries *all* the information. Hence the failure to identify any real movement for Labour from 1983 to 1987 to 1992 must stem from more than a mistaken methodology, like the very strong move by the Liberals 1992–1997 (but in the wrong direction).

BL and ourselves agree that the results in Figs. 1 and 2 of Budge and Pennings (2007) are disappointing. Whatever the reason, it rules out the most obvious way to get an authoritative basis for policy estimates based on word frequencies, which is bad news for them as promoters of the approach.

2. Methodological aspects of the Manifesto (MRG/CMP) data

We could have arrived at the essentially contested nature of word frequency estimates simply on the basis of the expert scores (Castles and Mair, 1984) and general judgements of party specialists about British and American party movements. The Manifesto estimates

[4] In spite of claims that the major application is cross-sectional, five out of six of the tests reported for Wordscores were time based (Laver et al., 2003, 319-328).

refine these but also conform to them (Budge et al., 2001, 24–26).

It is ironic therefore that BL devote two-thirds of their space to criticizing the status of these data as a 'benchmark'. In any case, even if they see some flaws in them, why not accept them as the major data set around and check results against them 'as if' they were the standard for the time being? (following McDonald and Mendes (2001) in comparing the Manifesto estimates with expert ones: see also McDonald et al. (2007)).

However, their critique, though misplaced in this context, does provide an opportunity for dealing with misconceptions which also occur in their earlier piece in the special issue. These often spring from their implicit assumption that the Manifesto project is like a finite expert or other survey, or a given word-frequency analysis,[5] and ought therefore to conform to procedures and checks appropriate to them. So it is worth emphasising the aspects which set it apart from practically all other policy estimates in the field:

- The most obvious is the open-ended nature of the data set, which is constantly expanding to cover new countries (at present 55) and new elections (up to 2005). This means that checks which were valid and comprehensive when they were done become outdated as they no longer relate to the new set five or ten years on. No single neat statistic can therefore be cited to cover all the data. The alternative is to publish intensive reviews of their validity and reliability at certain points in time (ignored by BL but see Budge et al., 2001, 111–141[6]). These efforts will be updated for the set extended to Central and Eastern Europe 1990–2003 (Klingemann et al., 2006, in press).
- To secure comparability and reliability a heavy emphasis is put on uniformity of procedures and central supervision and control. These are also extensively documented (Volkens, this issue); see also the massive Appendices in Budge et al. (2001), Klingemann et al. (2006). The procedures make micro level (inter-coder) checks uninformative in terms of the general validity and reliability of the data set as the operation is collective not individual.
- This collective nature of the project also sets it apart. The Manifesto Research Group has always taken decisions collectively, through regular meetings, correspondence, position papers, memos, detailed minutes of meetings, etc. This renders BL's footnote 3 plain wrong: the MRG Left-Right scale came out of collective discussion and meetings in 1986–1987 and Laver and Budge (1992) simply summarized their results in the introductory chapters (which is not to say they did not, like other participants, contribute to the discussions and work of the group).

Against this general background we now take up BL's criticisms systematically.

2.1. *'There is no method for assessing the uncertainty of the CMP estimates …' (p. 1)*

Had BL said 'there is no one method' we would agree in light of our first point above. What is certainly clear is that test scores for coders in (re-)training (unsupervised coding of a manifesto in English, often by a non-native speaker who is going to do production coding in their own language, and who will be given further training and supervision if they fail the test) are irrelevant to an assessment of the finished data set. They are also five times removed from the highly aggregated Left-Right scale used with Wordscores.

What is particularly surprising is that BL choose not to quote the published reliability statistic for the Left-Right time series (Heise, 1969), which is surely bang-on for their purpose and also appeared in a book edited by Laver (2001, 106) (see also Budge et al., 2001, 139). It is 0.942 for three time points in the 1980s and 1990s, which shows the estimates are highly reliable, and give useful information and only limited error in terms of variation. A forthcoming publication (Klingemann et al., 2006) will provide further error estimates on the basis of the Heise (1969) measurement model and other models. Error estimates can of course be calculated as the complement of Heise reliability coefficients for any set of three party-panel time series by anyone who is interested, on the basis of the published data.

It is of course individual scores that are in question here. Most MRG projects involve data aggregation through factor analysis or investigating relationships

[5] One concrete example of this is when they suggest that manual coding is more reliable when the text is longer (BL, Table 1). This would be true for word-frequencies of course. But it is probably reversed for a human coder who would find a short, cogent text easier to code than a long rambling one.

[6] A particularly interesting study was Hearl's (2001, 111–125) detailed comparison of results obtained from analysing the data set as the MRG left it (1946–1983) and from analyzing the data as extended up to 1996. This really gets at the question of 'moving benchmarks' as BL (Section 3) put it. The stability of the data structures reassures us about this. However, the check is not a standard one and cannot be summarized in a few statistics and is therefore ignored.

between variables through standard tests. Both of these make inbuilt estimates of error, so it is not 'unknown' in a general sense. Left-Right and other policy-specific indicators have been used to post dict other policy series and they work well (McDonald et al., 1999), which must surely increase general confidence in them.

Given the limits on individual coding checks mentioned above the MRG/CMP have favoured macro-level assessments of error because their broad range is better adapted to a general 'take' on the changing data set, and also because users want measures directly related to the actual variables they are using. Just after the coding frame was created (1978–1979) members of the MRG were understandably anxious about the instructions being intelligible and uniformly applied. Many therefore carried out check-codes on their national data set with satisfactory results (Budge et al., 1987, 23-24, 48, 78, 119, 140, 183, 237, 276, 332, 355, 375). German data 'at the lower scale of tolerance' (301–302) was as a result recoded in the early 1990s.

This example illustrates how check coding forms part of the coding process rather than providing error estimates for the data. How could one generalize to today's data set from statistics produced around 1980, by different coders from today's, on the first 30 years of what are now 50 year time-series? This is a problem which critics used to time bound data have not really thought through. They are certainly better bases for assessment than training tests however (BL, Section 1).

2.2. *'The problem of moving benchmarks …'*

(a) BL note that 'the political lexicon changes over time' and this renders the coding scheme 'designated in the early 1980s' time bound. The suggestion is a perfectly legitimate one but again a generalization from word-frequency analysis. Clearly if words are taken as data and the vocabulary changes over time it is going to be difficult to create long time series by counting them. This is one area where manual coding has an advantage however for with a set of reasonably general categories (e.g. Social Justice) a human can code many physically different words into the category and thus cope with changing vocabulary over time.

This was demonstrated by Budge and Farlie (1977, 422) coding British manifestos and US platforms from 1922 to 1976 inclusive, periods when the political vocabulary undoubtedly changed. Their 35 categories were the core of the 56 MRG ones, which expanded as a result of specialists' demands for more specific categories to cover 'their' country, and for 'pro' and 'con' positions on matters like the military. Many of the additional categories were thinly populated and created 'noise' (Laver and Budge, 1992, 23). However they can be aggregated to form a set of broader categories which eliminate most of this and bear a strong resemblance to those used by Budge and Farlie (1977) for 1922–1976.

Redesigning the scheme today would certainly involve this kind of aggregation but then it also did in 1986. Signs that the coding scheme was seriously time bound would be increases in uncoded sentences over time, which have not occurred. It has also been successfully used under the very different political conditions of Central and Eastern Europe after 1990 (Klingemann et al., 2006). Thus the scheme appears sufficiently time-invariant to cope with the post-war period so far, and certainly with 1979–1997 in Figs. 1 and 2 of Budge and Pennings (2007).

(b) BL state that the Left-Right scale is an 'inductive product … of coding categories used to define (it) … that loaded together in country-based exploratory factor analyses (for) 1945–1985 (bracketed words ours). This ignores the fact that 'a priori theoretical coherence was the prime consideration' (Laver and Budge, 1992, 26) in creating the scale. The reason categories in the coding frame were put together by the MRG was that they had *already* been put together by ideologues at the end of the 19th century. Marxists argued that State economic and social intervention were necessary to protect the working class against an exploitative system which propped itself up by overseas imperialism. Right-wing theorists (e.g. Green and his followers) argued the traditional order guarantees security and (market) freedom for everybody at home and abroad. If the Left-Right scale is time-bound therefore it is tied to 1900 and the founding party ideologies rather than 'centred on 1965'. It is theoretically coherent to the extent these ideologies are coherent and it has a clear content. One cannot query the rationale of the MRG scale in operationalizing this: it has a more coherent justification than most scales. One may want to decompose it into constituent parts (economic, social. foreign, etc.) for other purposes but it can hardly be criticised per se for following through on the ideological arguments. These can be and have been operationalized in terms of inductive coding categories opposed to each other by the theoretical reasoning. Policy positions are measured

by simple addition and subtraction, no category being theoretically privileged above the others.

Political scientists are used to thinking of scales in factor analytic terms, as inductively derived, with the association of categories being contingent on recent political developments. In this context the MRG scale is criticised for not including 'environment and immigration', or for being a priori as opposed to Gabel and Huber's inductive approach (BL, end Section 2). But this is the whole point. New issues are not part of the scale as conceived in terms of 1900. It needs to be kept pure and tied to classic arguments *in order* to be time invariant.

Is it then relevant to the post-war period? That, of course, is an empirical question: we can see if it is relevant by using it to estimate party positions and seeing if they make sense in given periods like the earlier or later post-war.

This was what the exploratory factor analyses of 1986 referred to by BL were intended to do—not to create the scale already formed a priori but to check its applicability. As the MRG wanted at the time to measure policy distances to relate to coalition formation, it was naturally concerned that the designated categories should hang together under factor analysis of the post-war data and produce intelligible 'maps' of party movement. The group would certainly have gone on to create alternative inductive measures had it not 'worked' in this way. But it did and has gone on doing so in a variety of countries up to now (Budge et al., 2001; Klingemann et al., 2006, Chapter 1).

As usual BL's critique is that the MRG scale is not structured like other attempts at measuring Left-Right. We would largely agree with their comments on the other attempts—lack of clear content, failure to keep up with new issues or incoherence in absorbing them. Hopefully, this discussion demonstrates that the MRG scale *has* clear content, hangs together coherently, is time invariant and—so far—continues to be empirically relevant.

3. Summary overview

This reply has demonstrated that Wordscores is not the only word frequency approach we discussed in our article (Kleinnijenhuis and Pennings were central as well) but that it shares in the weakness of all such approaches—what documents can authoritatively serve as calibrating/reference texts for the others?—and if absent, how stable and reliable will policy estimates be? We do however agree that much work remains to be done both on word frequencies and the Manifesto data. In the latter case, it will be reported in a forthcoming volume (Mapping Policy Preferences II, Klingemann et al., 2006). In the former case, new opportunities for computerized content analysis are created by the newly established digitalised collection of party manifestos (Pennings and Keman, 2002).[7]

References

Benoit, K., Laver, M., 2007a. Benchmarks for text analysis: a response to Budge and Pennings. Electoral Studies 26 (1), 130–135.

Benoit, K., Laver, M., 2007b. Estimating party policy positions: comparing expert surveys and hand-coded content analysis. Electoral Studies 26 (1), 90–107.

Budge, I., Farlie, D.J., 1977. Voting and Party Competition. Wiley, New York.

Budge, I., Robertson, D., Hearl, D.J. (Eds.), 1987. Ideology, Strategy and Party Change. Cambridge University Press.

Budge, I., Klingemann, H.-D., Volkens, A., Bara, J., Tanenbaum, E., et al. (Eds.), 2001. Mapping Policy Preferences: 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.

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

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.

Hearl, D.J., 2001. Checking the party policy estimates: reliability. In: Budge, et al. (Eds.), Mapping Policy Preferences: Estimates for Parties, Electors and Governments 1945–1998. Oxford University Press, Oxford, pp. 111–126.

Heise, D.R., 1969. Separating reliability and stability in test-retest correlation. American Sociological Review 33, 93–101.

Kleinnijenhuis, J., Pennings, P., 1999. A probabilistic keyword approach to textual analysis. Paper presented at Mannheim Joint Sessions of the ECPR.

Kleinnijenhuis, J., Pennings, P., 2001. Measurement of party positions on the basis of party programmes, media coverage and voter perceptions. In: Laver, M. (Ed.), Estimating the Policy Position of Political Actors. Routledge, London, pp. 162–182.

Klingemann, H.-D., Volkens, A., Bara, J., Budge, I., McDonald, M., 2006. Mapping Policy Preferences II: Estimates for Parties, Electors and Governments in the OECD, EU and CEE 1990–2003. Oxford University Press, Oxford.

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

Laver, M. (Ed.), 2001. Estimating the Policy Position of Political Actors. Routledge, London.

[7] The digitalised party manifestos are made available upon request by the Zentralarchiv für Empirische Sozialforschung, Universität zu Köln, under certain conditions regarding their usage. The data are part of the Comparative Electronic Manifestos Project (http://research.fws.vu.nl/DoingResearch) directed by Paul Pennings and Hans Keman, Vrije Universiteit Amsterdam, The Netherlands Organisation for Scientific Research (project # 480-42-005).

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. (Eds.), 1992. Party Policy and Government Coalitions. St. Martins Press, New York.

McDonald, M.D., Budge, I., Hofferbert, R.I., 1999. Party mandate theory and time series analysis. Electoral Studies 18, 587–596.

McDonald, M.D., Mendes, S.M., Kim, M., 2007. Cross temporal and cross-national comparisons of party left-right positions. Electoral Studies 26 (1), 62–75.

McDonald, M., Mendes, S.M., 2001. Checking the party policy estimates: convergent validity. In: Budge, et al. (Eds.), Mapping Policy Preferences: Estimates for Parties, Electors and Governments 1945–1998. Oxford University Press, Oxford, pp. 127–142.

Pennings, P., Keman, H., 2002. Towards a new methodology of estimating party policy positions. Quality and Quantity 36, 55–79.

Volkens, A., 2007. Strengths and weaknesses of approaches to measuring policy positions of parties. Electoral Studies 26 (1), 108–120.