B.J.Pol.S., Page 1 of 15 *Copyright © Cambridge University Press, 2012*
doi:10.1017/S0007123412000105

Notes and Comments

Natural Sentences as Valid Units for Coded Political Texts

THOMAS DÄUBLER, KENNETH BENOIT, SLAVA MIKHAYLOV AND MICHAEL LAVER*

A rapidly growing area in political science has focused on perfecting techniques to treat political text as 'data', usually for the purposes of estimating latent traits such as left–right political policy positions.[1] More traditional approaches have applied classical content analysis to categorize sub-units of political text, such as sentences in manifestos. Prominent examples of this latter approach include the thirty-year old Comparative Manifestos Project and the Policy Agendas Project.[2] 'Text as data' approaches use machines to convert text to quantitative information and use statistical tools to make inferences about characteristics of the author of the text. Content analysis schemes use humans to read textual sub-units and assign these to pre-defined categories. Both methods require the prior identification of a textual unit of analysis – a highly consequential, yet often unquestioned, feature of research design.

Our objective in this Research Note is to question the dominant approach to unitizing political texts prior to human coding. This is to parse texts into *quasi-sentences* (QSs), where a QS is defined as part or all of a natural sentence that states a distinct policy proposition. The use of the QS rather than a natural language unit (such as a sentence defined by punctuation) is motivated by the desire to capture all relevant political information, regardless of the stylistic decisions made by the author, for example, to use long or short natural sentences. The identification of QSs by human coders, however, is highly unreliable. If, comparing codings of the same texts using quasi-sentences and natural sentences, there is no appreciable difference in measured political content, then there is a strong case for replacing 'endogenous' human unitization with 'exogenous' unitization based on

* Däubler (email: thomas.daeubler@mzes.uni-mannheim.de) is at the Mannheim Centre for European Social Research (MZES); Benoit is at the London School of Economics (Methodology Institute) and Trinity College Dublin (Political Science); Mikhaylov is at University College London (Department of Political Science), and Laver is at New York University (Department of Politics). The authors would like to thank Daniela Braun and Hermann Schmitt for valuable comments and kind permission to reproduce some of the experimental results discussed in Braun, Mikhaylov and Schmitt (2010). An earlier version of this Research Note was presented at the ECPR Joint Sessions of Workshops in St. Gallen, 2011, and the authors would also like to thank the participants in the workshop entitled 'The how and why of party manifestos in new and established democracies', for helpful feedback.

[1] For example Michael Laver, Kenneth Benoit and John Garry, 'Extracting Policy Positions from Texts Using Words as Data', *American Political Science Review*, 97 (2003), 311–31; Jonathan Slapin and Sven-Oliver Proksch, 'A Scaling Model for Estimating Time-Series Party Positions from Texts', *American Journal of Political Science*, 52 (2008), 705–22.

[2] Ian Budge, Hans-Dieter Klingemann, Andrea Volkens, Judith Bara, Eric Tanenbaum, Richard C. Fording, Derek J. Hearl, Hee Min Kim, Michael D. McDonald and Silvia M. Mendes, *Mapping Policy Preferences: Estimates for Parties, Electors, and Governments, 1945–1998* (Oxford: Oxford University Press, 2001); Hans-Dieter Klingemann, Andrea Volkens, Judith Bara, Ian Budge and Michael McDonald, *Mapping Policy Preferences II : Estimates for Parties, Electors, and Governments in Eastern Europe, European Union, and OECD 1990–2003* (Oxford: Oxford University Press, 2006); Frank R. Baumgartner, Christoffer Green-Pedersen and Bryan D. Jones, *Comparative Studies of Policy Agendas* (London: Routledge, 2007).

natural sentences that can be identified with perfect reliability by machines using pre-specified punctuation delimiters.

We proceed as follows. First, we discuss the issues motivating the use of QSs and the implications of this for reliability. Next, we re-examine and recode, using natural sentences, a set of texts in several languages that have previously been unitized and coded using QSs, comparing results to see if this generates significant differences in measured political content. We also compare coding reliabilities that arise when using natural sentences rather than quasi-sentences. Our results provide strong evidence that unitizing text exogenously using natural sentences is systematically more reliable than endogenous unitization based on human judgement, while delivering substantively similar estimates of the positions of text authors.

THE RATIONALE FOR ENDOGENOUSLY DEFINED TEXT UNITS

Expert text coding involves two data-generating steps: dividing text into units of analysis, then assigning each text unit a code.[3] While this second step (coding) typically receives the greatest scrutiny, our concern here focuses on the first step (unitization). Prior to coding, a text must be *unitized* by dividing it into smaller units relevant to the research question. Unitization can be specified *exogenously* to the research process using no human judgement, on the basis of predefined rules. This defines units of text in a manner independent of any coding decisions made as part of the analysis. Examples of such rules include using words, word sequences or *n*-grams, natural sentences, paragraphs, pages or even entire documents as the unit of analysis. Alternatively, text units may be defined *endogenously* to the coding process and involve human judgement, *as part of the content analysis itself*, to determine where one unit of content ends and another begins.[4] Artificial intelligence applications of natural language processing identify text units during the text processing procedure, as do the QS parsing schemes applied by human coders in the Comparative Manifestos and Policy Agendas Projects. Choosing whether to define text units exogenously or endogenously to the text coding procedure involves making a fundamental trade-off between the goals of reliability and validity.

Expert or 'hand' coding methods are not alone in facing the issue of how to define the unit of text analysis. *Statistical scaling or classification methods*, in which there have been numerous recent advances,[5] typically make the linguistic 'bag of words' assumption and specify the word as the fundamental unit of text analysis. All substantive decisions about text unitization are part of the research design, not the coding process. *Dictionary-based methods* apply a pre-defined coding dictionary, the substantive content of which is at the heart of the research design, to tag words or word stems with coding categories associated with these words by the dictionary. As with statistical methods, the goal is typically fully automated machine coding, with all substantive decisions made as part of the research design rather than the unitization or coding processes. In methods of *automated natural language processing,* the goal is the automated extraction of meaning from natural language. Well-known

[3] Klaus Krippendorff, *Content Analysis: An Introduction to its Methodology* (Thousand Oaks, Calif.: Sage, 2004).

[4] By contrast, the second basic data-generating step, in which each text unit is coded by assigning to it a category from the coding scheme, is always endogenous to the text, and indeed forms the core part of the content analysis exercise.

[5] Laver, Benoit and Garry, 'Extracting Policy Positions from Texts Using Words as Data'; Burt L. Monroe and Ko Maeda, *Rhetorical Ideal Point Estimation: Mapping Legislative Speech* (Palo Alto, Calif.: Stanford University: Society for Political Methodology, 2004); Burt Monroe, Michael Colaresi and Kevin M. Quinn, 'Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict', *Political Analysis*, 16 (2008), 372–403; Slapin and Proksch, 'A Scaling Model for Estimating Time-Series Party Positions from Texts'; Daniel J. Hopkins and Gary King, 'A Method of Automated Nonparametric Content Analysis for Social Science', *American Journal of Political Science*, 54 (2009), 229–47.

examples can be found in Google Translate, or the Watson system recently and successfully developed by IBM, part of the 'DeepQA' project to understand and then answer the complex natural language questions that form part of the *Jeopardy* television quiz programme.[6] In such applications, the unit of text analysis must be interpreted *as part of the system that also processes the text*, rather than exogenous to the text analysis. Thus far, this research programme has been the preserve of computer scientists and computational linguists, as opposed to political scientists.

The method deployed by research projects based on text coding by human experts, such as the Comparative Manifestos and Comparative Policy Agendas Projects, is in essence non-automated natural language processing. Crudely speaking, expert coding can be seen as using skilled humans to engage in complex pattern recognition tasks that we cannot yet programme computers to produce with valid results. As with all natural language processing, the fundamental unit of text analysis may transcend punctuation marks, may conceivably range from a short phrase to an entire text corpus, and may also require human decisions about where a relevant text unit begins and ends. Because only humans can yet be trusted to provide *valid* readings of complex texts for meaning, such traditional methods of content analysis inherently involve subjective judgements by humans reading, parsing and coding text. Introducing human judgement as part of the text coding process rather than just the research design, however, introduces serious concerns about reliability that are not an issue for the other methods of text analysis we have identified.

RELIABILITY TRADE-OFFS WITH SUBJECTIVE TEXT UNITIZATION

The core issue when designing schemes for unitizing and then coding text units as part of an expert coding project concerns the classic trade-off between reliability and validity. In expert text coding, a procedure is reliable when 'the reading of textual data as well as of the research results is replicable elsewhere, [so] that researchers demonstrably agree on what they are talking about'.[7] Whenever non-deterministic instruments – such as human beings – are used to unitize and code texts, then the content analysis procedure faces potential problems with reliability. No matter how carefully trained, human coders invariably disagree over qualitative decisions regarding the identification of text units and their classification into coding categories. Depending on how unreliable the procedure is, estimates constructed from the codings may also lack *validity* because of the level of noise or even bias introduced by the content analysis procedure. Reliability is no guarantee of validity, however, and in practice validity tends to suffer in the pursuit of maximizing reliability. Indeed, the debate over machine versus human coded content analysis largely revolves around the desire to balance the competing objectives of reliability and validity. Proponents of computerized schemes for estimating party positions from political manifestos cite perfect reliability in their favour, yet struggle to demonstrate validity.[8] Hand-coded schemes such as the CMP claim validity as a central advantage but then devote huge resources to attempts to enhance reliability.[9]

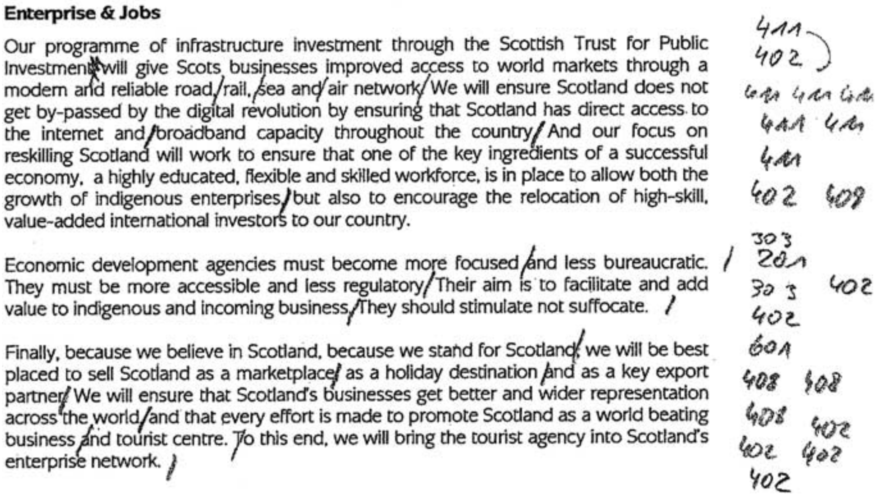
Validity, at its simplest, means that the results of some content analysis can be seen to reflect the 'true' content of the text in a meaningful way. If researchers use expert coders to classify the content of texts then, these days, they almost surely have chosen this laborious method over machine coding because they feel that the results are more valid – that humans can currently extract more valid meaning from complex texts than machines can. Furthermore, expert codings are likely to be most valid when the unit of text analysis is endogenous, since it is unlikely that readers

[6] Details for those unfamiliar with the *Jeopardy* game show format, or with Watson and DeepQA, can be found at http://www-03.ibm.com/innovation/us/watson/.

[7] Krippendorff, *Content Analysis: An Introduction to its Methodology*, p. 212.

[8] Michael Laver and John Garry, 'Estimating Policy Positions from Political Texts', *American Journal of Political Science*, 44 (2000), 619–34; Laver, Benoit and Garry, 'Extracting Policy Positions from Texts Using Words as Data'; Slapin and Proksch, 'A Scaling Model for Estimating Time-Series Party Positions from Texts'.

[9] See, for instance, Klingemann *et al.*, *Mapping Policy Preferences II*, chaps 4–6.



Enterprise & Jobs

Our programme of infrastructure investment through the Scottish Trust for Public Investment/will give Scots businesses improved access to world markets through a modern and reliable road/rail,/sea and/air network/We will ensure Scotland does not get by-passed by the digital revolution by ensuring that Scotland has direct access to the intemet and/broadband capacity throughout the country/And our focus on reskilling Scotland will work to ensure that one of the key ingredients of a successful economy, a highly educated, flexible and skilled workforce, is in place to allow both the growth of indigenous enterprises/but also to encourage the relocation of high-skill, value-added international investors to our country.

Economic development agencies must become more focused/and less bureaucratic. / They must be more accessible and less regulatory/Their aim is to facilitate and add value to indigenous and incoming business/They should stimulate not suffocate. /

Finally, because we believe in Scotland, because we stand for Scotland/we will be best placed to sell Scotland as a marketplace/as a holiday destination/and as a key export partner/We will ensure that Scotland's businesses get better and wider representation across the world/and that every effort is made to promote Scotland as a world beating business/and tourist centre. /To this end, we will bring the tourist agency into Scotland's enterprise network. /

411
402
411 411 411
411 411
411
402 409
303
201
303 402
402
601
408 408
408
402
402 402
402

Fig. 1. Section of SNP 2001 manifesto parsed into quasi-sentences by CMP coder

for whom the texts are written pay close attention to punctuation marks when they read a text for meaning. Despite the ongoing debate about whether automatic coding can produce valid representations of textual content, our focus here applies only to the validity versus reliability issue applied to unitization. If we can demonstrate that deterministic unitization rules – axiomatically meeting a perfect standard of reliability in unitization – are just as valid as subjective ones based on endogenously defined text units, then we can move the state of human-coding text research further out on the reliability–validity frontier by eliminating the unreliability of subjective unitization without suffering a trade-off in validity.

The two most widely used coding schemes in political science – the Comparative Manifestos Project and the Policy Agendas Project[10] – both specify the unit of textual analysis as an endogenous text fragment known as the *quasi-sentence* (QS): 'an argument which is the verbal expression of one political idea or issue'.[11] This approach to unitizing is often referred to as thematic unitizing.[12] The explicit motivation for using QSs is to avoid missing separate policy statements from political texts created by more long-winded authors who tend to combine multiple policy statements into single natural sentences. More generally, the rationale for endogenous text unitization is to implement a method of natural language processing when the meaning in natural language may not respect punctuation marks.

As an example, consider this natural sentence taken from the 2001 Australian National Party manifesto. The tenth natural sentence states: 'We know that the only way to create economic prosperity is to rely on individual enterprise / and we know that our future as a nation depends on having strong families and communities.' The CMP coder of this document identified two QSs, indicated here by the '/'. The first was assigned to category 401 (Free Enterprise: Positive), the second to category 606 (Social Harmony: Positive). To see how quasi-sentences' unitization is executed in practice on a somewhat larger scale, in Figure 1 we have reproduced a section of the Scottish National Party manifesto of 2001 – parsed into QSs by a coder from the Comparative Manifestos Project. QSs are demarcated by the pencil marks in the text, indicating that ten natural sentences have been divided into twenty-three

[10] Budge *et al.*, *Mapping Policy Preferences*; Baumgartner, Green-Pedersen and Jones, *Comparative Studies of Policy Agendas*.

[11] Andrea Volkens, *Manifesto Coding Instructions* (2nd revised edn), Wissenschaftszentrum Berlin, Discussion Paper FS III 02-201 (2001), p. 96.

[12] Krippendorff, *Content Analysis*.

QSs, some as short as a single word. It is clearly not self-evident that different coders would meet Krippendorff's – or indeed anyone's – definition of reliability given the large number of different, and perfectly reasonable, ways of identifying an alternative set of word strings qualifying as independent QSs from this short manifesto fragment.[13]

Using carefully trained expert coders following well-defined instructions may mitigate problems of unitization unreliability, but we now show that serious problems remain. We obtained data from the CMP, derived from their own coder training experiments, in which many expert coders were asked to unitize and code the same training document, and the results from sixty-seven trained coders showed huge variability in the number of QSs they identified in the text. According to the CMP's master coding, applying the authoritative version of their quasi-sentence unitization, the document contains a 'true' number of 163 QSs. The CMP's trained expert coders, however, identified a total number of QSs ranging from about 120 to 220, with a standard deviation of 19 and an interquartile range of 148 to 173.[14] The conclusion is clear. When even well-trained human expert coders specify units of analysis endogenously – as CMP coders do when they parse a text into QSs – the results are very unreliable. Because human-coded content analyses typically combine results from different coders, furthermore, systematic differences in subjective judgements about endogenously identified text units (whereby some coders tend to see more units where others tend to see fewer) may introduce bias as well as unreliability.

One way to enhance the reliability of expert text coding is to develop an automated method for *endogenously* identifying QSs. The hard problems of natural language processing for complex political texts, however, mean that no automated unitization of texts into QSs is currently feasible – at least, none which we would confidently declare is valid for making the information-rich thematic distinctions motivating the use of endogenous text units in the first place. An alternative approach is to define text units *exogenously* to the content analysis process, following (for example) syntactical distinctions that are 'natural' relative to the grammar of the text.[15] Among the choices of syntactically delimited units, *natural sentences* are closest to the thematically defined QSs used by the CMP. Instead of endogenously defined thematic units, natural sentences are exogenously specified using predefined lists of punctuation marks. The open empirical question, addressed in the rest of this Research Note, concerns whether specifying the unit of analysis as exogenously specified natural sentences, rather than endogenously specified QSs, significantly affects inferences about the substantive content of the types of text we wish to investigate.[16] Exogenous specification of the unit of text analysis as a natural sentence is axiomatically more reliable than allowing expert coders to unitize text endogenously. If exogenous unitization does generate different results, this raises the reliability–validity trade-off for consideration. If it does not, then using perfectly reliable natural sentences as the fundamental unit of text analysis for expert coding is a dominant methodological strategy.

[13] To return to the Australian 2001 National party example cited earlier, we also observe the following natural sentence: 'There is no argument about the need for production sustainability and its matching twin, environmental sustainability.' In this case, the coder deemed this a single QS and coded it as 501 (Environmental Protection: Positive), even though it could plausibly have been seen as comprising two QSs, divided by the 'and', with the first coded to 410 (Productivity: Positive) and the second to 501.

[14] In the test results, coders with especially bad first round results had these corrected, and were asked to repeat the experiment. Here we report only the second-round unitization results for coders asked to repeat the test. While these results are not a decisive experiment, given that it is part of a training process of new coders, they are the single largest test of multiple unitizations of a manifesto text available. We thank Andrea Volkens for sharing this data with us.

[15] Krippendorff, *Content Analysis*.

[16] The use of natural sentences for manifesto coding is also mentioned by Leonard Ray, 'A Natural Sentences Approach to the Computer Coding of Party Manifestos', in Michael Laver, ed., *Estimating the Policy Positions of Political Actors* (New York: Routledge, 2001), pp. 149–61. The discussion there, however, makes no clear distinction between natural and quasi-sentences, and for the most part describes possible future approaches to automated natural language processing of political texts.

DATA AND METHODS

Our comparison of the validity of expert-coded text analyses based on exogenous versus endogenous text units comes from a reanalysis of manifestos originally unitized and coded by the CMP. Ideally, we would provide a set of manifestos to a large group of coders, and ask that each be coded on the basis of natural sentences and QSs, and then compare the aggregate measures of political content. If there were no appreciable differences in the measures of aggregate political content, then we would declare both methods equally valid.[17] Of course, this comparison would not determine whether either method in itself was valid in absolute terms, but if no differences exist in the way each unitization scheme characterizes political content, then it is strong evidence that one cannot be considered less valid than the other.

Without conducting a time-consuming and expensive test such as this, however, we do have access to information that allows us to investigate whether endogenous text unitization makes a difference. This involves returning to a set of manifestos previously unitized into QSs and coded by trained CMP coders, codings that form part of the data reported in the CMP dataset. This set of fifteen documents consists of printed manifestos with unitization marks and marginal codes of the sort depicted in Figure 1. From a limited number of such texts that we were able to obtain upon request from CMP archives, we selected fifteen texts with the aim of incorporating a large range of political contexts. We maximized the number of countries covered and included a wide variety of texts in terms of language, length, party family of the authoring party and left–right orientation. The sample includes eight English texts, consisting of one manifesto from Australia, one from Ireland, one from New Zealand, three from Britain and two from the United States; three Estonian manifestos (in Estonian); two German-language manifestos from Austria; and two manifestos from Iceland (in Icelandic). A more detailed description of the sample is provided in Appendix Table A1.

Using the selected texts, we proceeded in two steps. First, we recorded all QS codes indicated on the margin of the documents, and also noted whether each coded QS is identical to a natural sentence or a fragment of some natural sentence. This yields a dataset where the unit of the analysis is the natural sentence, while component QSs (one or several) are sub-units.[18] In the second step, we assigned a CMP policy code to each natural sentence. In this step, three different situations can occur:

1. *A natural sentence contains one single QS*. In this case, the CMP policy code assigned to the natural sentence is simply the one originally assigned to the QS.
2. *A natural sentence contains more than one QS, but all were given the same policy code by the CMP coder*. In this case, the natural sentence receives this code.
3. *A natural sentence contains more than one QS, and these were given different policy codes*. In this case, a human coding the natural sentence and faced with this choice would probably decide which of the competing policy codes best represented the natural sentence unit. Our procedure applied three different rules for making such a coding decision:
   First: Assign the natural sentence the code of the first component QS.
   Last: Assign the natural sentence the code of the last component QS.
   Random: Assign the natural sentence the code of a randomly chosen component QS.

To be as sure as possible of the validity of our recoding exercise, the authors themselves carefully applied this method to the fifteen selected manifestos.

[17] We are assuming here that differences at the unit level are not the quantity of interest, and that the objective of any unit-based coding exercise is to yield aggregate measures of political content.

[18] To make the identification of natural sentences as unambiguous as possible, with a view to eventually automating this stage completely, we developed a very explicit set of guidelines as to how to identify a natural sentence. A natural sentence delimiter was defined as the following characters: '.', '?', '!', and ';'. Bullet-pointed sentence fragments were also defined to be 'natural' sentences, even if not ending in one of the five previously declared delimiters. A full set of the coding instructions we issued to coders (ourselves) is available upon request.

TABLE 1 *Pattern of Natural Sentences versus Quasi-Sentences from 15 Election Manifestos*

Extent to which natural sentence was split into several QSs	Natural sentences (*N*)	Natural sentences (%)	% of natural sentences with different CMP codes	% of natural sentences with different left–right codes (left, neutral, right)
Not split	5,847	84.0	–	–
Split into two	832	12.0	44.1%	28.7%
Split into three	202	2.9	58.4%	36.6%
Split into four	50	0.7	76.0%	50.0%
Split into > 4	29	0.4	44.8%	31.0%
Total	6,960	100	7.7%	5.0%

We first report the frequencies of the three types of relationship between natural sentences and quasi-sentences. We then aggregate text codings into widely used left–right policy scales, and compare our substantive conclusions about document content when we shift from QS unitization to natural sentences (also comparing the *first*, *last* and *random* rules for assigning a QS code to the natural sentence level).

Human coders faced with a natural sentence that they feel contains more than one policy statement may face a tough choice in deciding which code to assign it, if only one code may be assigned.[19] In such a case, it is possible that *coding* reliability – a separate issue from unitization reliability – may become an issue. As a preliminary test of this possible problem, we report the results of a coding experiment conducted using an expanded version of the CMP scheme applied to European election manifestos, designed to test whether coding unreliability increased when coders were asked to use natural rather than quasi-sentences as the basis for the CMP scheme.

RESULTS FROM RECODING MANIFESTOS INTO NATURAL SENTENCES

Comparing Units of Analysis

Our revisiting and recording of the text units from the fifteen manifestos provided a dataset of a total of 6,960 natural sentences, in which were contained 8,481 QSs. These are described in Table 1.

The clearest result to emerge from our analysis is that the splitting by CMP coders of natural sentences into more than one QS occurs quite infrequently: in more than eight out of ten cases (84 per cent), natural sentences contained a single QS only, meaning that the endogenous unitization problem pertains to only 16 per cent of all text units. The remaining natural sentences include mostly two (12 per cent of all natural sentences) or three QSs (2.9 per cent of all natural sentences). Natural sentences with four or more QSs are very rare, making up just 1.1 per cent of our sample.

[19] We assume that only one category can be assigned to each text unit. This follows standard approaches in content analysis that were implemented in the coding protocols of the CMP (see Klingemann *et al.*, *Mapping Policy Preferences II*, Appendix II, 'Manifesto Coding Instructions'). In order to allow for reliable content coding, there are two main requirements for the categories of a coding scheme: they must be mutually exclusive and exhaustive (Krippendorff, *Content Analysis*, p. 132). The possibility of several categories applicable to each text unit violates the exclusiveness requirement. Mutual exclusiveness is related to the ability of coders to conceptualize the text unit they are reading clearly. Lack of mutual exclusiveness results in semantic confusion that leads to misclassification and biased results. This aspect has been appreciated by the CMP who developed detailed decision rules on handling the situations when more than one category seems to apply to each text unit (Klingemann *et al.*, *Mapping Policy Preferences II*, Appendix II). However, as shown in experimental results (Slava Mikhaylov, Michael Laver and Kenneth Benoit, 'Coder Reliability and Misclassification in the Human Coding of Party Manifestos', *Political Analysis,* 20 (2012), 78–91), semantic confusion and related misclassification result in low reliability of the data generation process in the CMP.

The second strong result from our analysis is that when natural sentences are split into component QSs, these components are not necessarily coded differently. In fact, in the category of natural sentences with two QSs, fewer than half (44.1 per cent) of the natural sentences have different component codes, rising to just over half (58.4 per cent) for natural sentences split into three QSs. More of those split into four or more QSs were different, although overall, as previously mentioned, these represent a tiny fraction of all of natural sentences. Considering all natural sentences, the share of cases with varying component codes is just 7.7 per cent. In other words, before any additional comparison, we expect results that are at the very least 92 per cent identical, because there is a 92 per cent similarity between the two unitizations. Coding all QSs in the same natural sentence into the same category is at odds with the rationale for using QSs in the first place, which is that a natural sentence may have more then one component of meaning. Indeed, it amounts to an arbitrary, and as we have seen unreliable, 'double' or 'triple' counting of text units in which one natural sentence contains several identically coded QSs.

Results reported so far refer to coding differences based on the 56-category CMP scheme.[20] Published applications of CMP data typically use scales that combine many coding categories. By a country mile, the most popular application of the CMP data is the left–right index 'Rile', a scale that combines twenty-six of the fifty-six CMP coding categories. It is quite possible that all QSs in the same natural sentence were coded in the same direction (left, right, or neither) in relation to the Rile scale. The fifth column of Table 1 shows that, among natural sentences with two component QSs, less than three out of ten (28.7 per cent) involve QSs coded in different directions on the Rile scale. Among natural sentences with three component QSs, this figure is similar (36.6 per cent). The total share of natural sentences with component QS codes that differ in terms of left–right orientation is only 5 per cent.

In a purely descriptive sense, our analysis comparing natural sentence to QS unitization has shown that, even prior to our comparison of substantive political content, we should expect similarities of 92 and 95 per cent between codings based on perfectly reliable exogenously defined text units, natural sentences, and those based on unreliable, labour-intensive endogenously defined QSs – in practice these units of analysis are exactly the same in twelve out of thirteen cases.

Comparing Aggregate Results

Individual sentence codings are of little substantive interest to end users of political content analysis datasets, and are not even reported. Instead, the CMP dataset contains only the percentages of text units coded into each policy category – the 'per' codes – as well as the total number of QSs recorded in the manifesto. Our aim in this section is, therefore, to compare aggregate category percentages from each manifesto when these are reconstructed from natural sentences and QSs.[21] This involved applying our three coding rules – choose the first QS code, the last, or one at random – to code the natural sentence. We see from Table 1 that this affects just about 8 per cent of all natural sentences.

Figure 2 shows the comparison of each policy category's percentage share, in a scatterplot matrix comparing our three rules for coding natural sentences with more than one QS, to the raw QS-based results used by the CMP. Each point represents a policy percentage from one manifesto, and the

[20] To be precise, the number of categories is 57 since it includes 'uncoded' as a further category, as is the case in the published CMP data. We did not use the four-digit codes that apply to post-communist countries but aggregated them to their respective three-digit category. However, this affected only 7 out of the total 8,481 (0.08%) QSs. In addition, 24 QS codes (0.28%) could not be identified from the documents since they were not legible.

[21] The emphasis here is on 'reconstructed': we did not ensure that every category percentage from the QS we recorded perfectly matched those reported in the CMP's dataset. An exact replication is not possible, for instance because it appears (not that rarely) that the number of codes on the margins does not correspond to the number of units separated by tick marks (if they are used at all). While we did check that we matched the published figures to a very high degree, a perfect matching is unnecessary since our comparison focuses on units within texts.

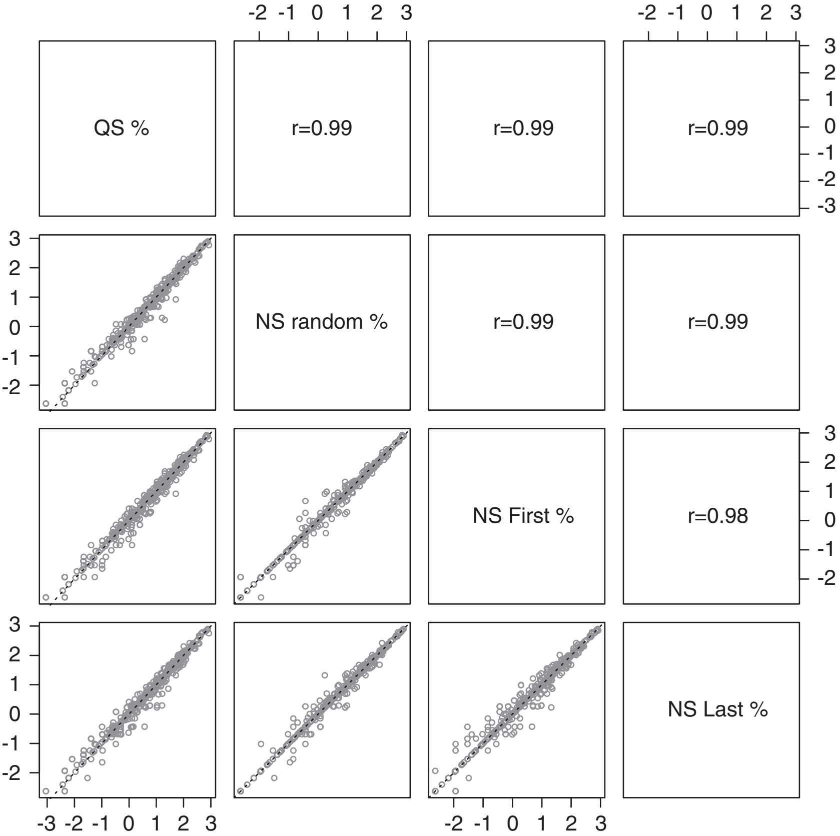


Fig. 2. Comparing quasi-sentence aggregate category percentages to natural sentence recodings
Notes: Three rules are compared: randomly assign the code based on constituent QSs; take the first QS code for the natural sentence; and take the last QS code for the natural sentences. Total manifestos analysed: 15.

dashed line shows the 45° axis of perfect agreement. To reduce skew created by low-frequency policy categories, we logged both axes (this makes no substantial difference to the results). The squares above the diagonal report Pearson's correlation coefficient, ranging from 0.98 to 0.99 – there is an almost perfect linear relationship regardless of which rule is applied. To test the overall agreement in a more numerical framework, we used a simple regression analysis of the logged QS policy category percentages on the logged natural sentence policy category percentages (see Table 2). The results confirm inferences drawn from viewing scatterplots; 98 per cent of the variance in the original QS coding is explained by the natural sentence codings, regardless which rule is applied. An *F*-test of whether the estimated slope coefficient differs from the 1.0 value implying perfect identity cannot reject this null hypothesis. This is strong evidence that the natural sentence and QS codings yield the same aggregate results.

As we noted above, the most commonly used product of the CMP dataset is the left–right 'Rile' index that includes twenty-six of the fifty-six CMP coding categories. Figure 3 plots Rile scores for our fifteen manifestos, using exogenous natural sentence and endogenous QS unitization, and shows a very high degree of agreement between the two. Because Figure 3 only has one aggregate data point representing each manifesto for which we recorded the text units, we re-sampled natural sentences drawing 100 samples of 100 natural sentences from each manifesto. Figure 4 reports results, plotting a total of 15 × 100 = 1,500 points representing the fifteen manifestos in our sample. The top panel of Figure 4 shows the CMP's additive, original Rile scale, while the bottom panel depicts the recently proposed aggregate logit Rile scale, a scale that has demonstrably better

TABLE 2 *Regression of (Log) Quasi-Sentence-Based % Categories by Manifesto on (Log) Natural Sentence-Based Estimates Using Three Rules*

	Dependent variable: Log (quasi-sentence per)		
	(1) Random QS Code	(2) QS Code Last	(3) First QS Code
Log(natural sentence per)	**1.000** (0.007)	**0.995** (0.007)	**0.999** (0.007)
N	460	463	459
R^2	**0.98**	**0.98**	**0.98**
p-value for F-test of H_0: $\beta = 1.0$	0.99	0.50	0.90

Notes: The constant was constrained to be zero. The F-test reported in the last line is a test of the null hypothesis that the slope coefficient is the identity value of 1.0.

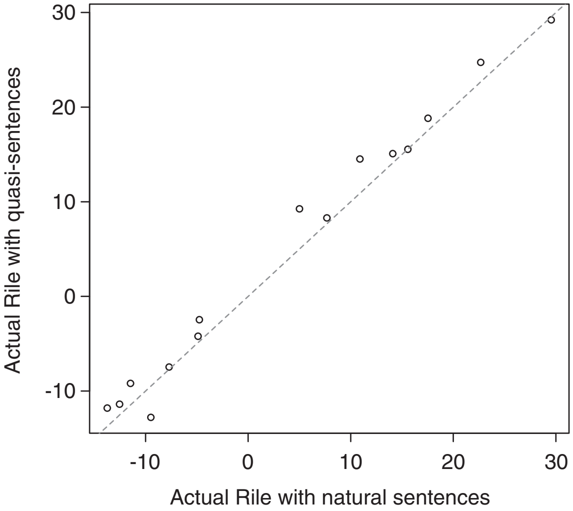


Fig. 3. Actual Rile Values aggregated for each manifesto
Note: Based on random assignment, which we have chosen because human coders could almost certainly do better than this rule.

properties than the CMP's relative difference scale.[22] In both cases, almost perfect correspondence is observed, even given the variation to be expected in each case from the sampling procedure.

RESULTS FROM A CODING RELIABILITY EXPERIMENT

The results we report above imply that using natural sentences rather than quasi-sentences as units of analysis does not affect the *validity* of the classification of these units following deterministic *unitization*. Indeed, we demonstrated that endogenous unitization so rarely results in multiple and

[22] Will Lowe, Kenneth Benoit, Slava Mikhaylov and Michael Laver, 'Scaling Policy Preferences from Coded Political Texts', *Legislative Studies Quarterly*, 36 (2011), 123–55. This index is constructed as $\log((R + 0.5)/(L + 0.5))$, where R and L are the summed percentages of the 13 right and left policy categories, respectively.

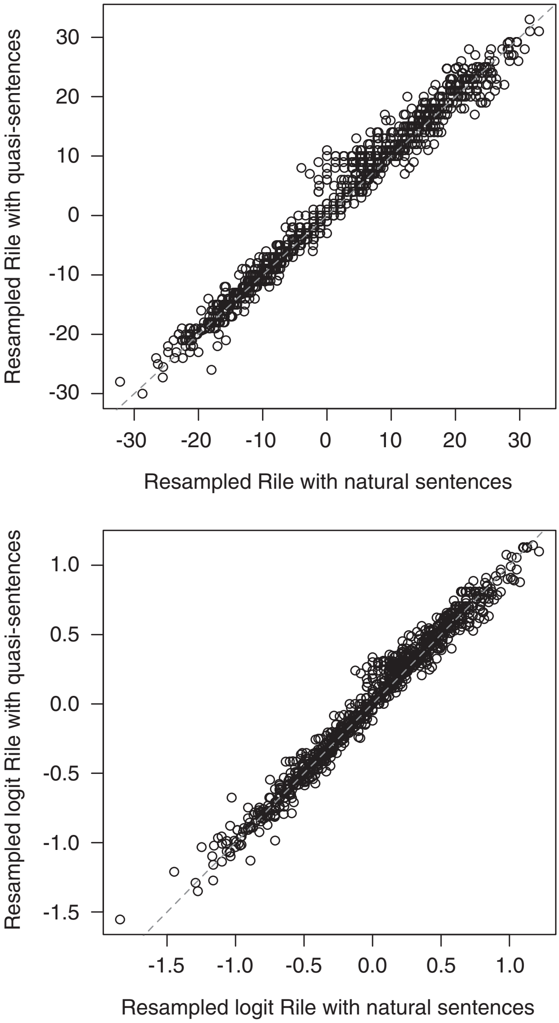


Fig. 4. Re-sampled Rile values
Notes: Based on random assignment, we took 100 random draws of 100 natural sentences each from each manifesto, and plotted the overall distribution of scores. The bottom plot uses the log Rile introduced by Lowe, Benoit, Mikhaylov and Laver, 'Scaling Policy Preferences from Coded Political Texts'.

differently coded QSs within one natural sentence unit that even random allocation of codes to the larger natural sentence units resulted in essentially the same aggregate results – suggesting that the reliability of *coding* has little potential to be adversely affected by the switch to natural sentence units. Our test used an exogenously specified procedure to code natural sentences, when these were split into differently coded QSs. A computer applying formal rules to do this will suffer no qualms of indecisiveness and display no favouritism towards particular policy categories or domains. It is possible, however, that a human expert coder faced with a natural sentence containing what are seen as two distinct policy statements will not use a consistent rule in coding the natural sentence text unit. While eliminating the unreliability of subjective unitization by using natural sentences, it could be that we are at the same time increasing the unreliability of coding by forcing human experts to make a Sophie's Choice when coding natural sentences that could, and perhaps should, be considered to express more than one distinct policy statement.

Our validity tests did not involve any human recoding, since the objective was to identify a lower bound of reliability by comparing three random allocation rules to code those 16 per cent of natural sentences that were split into multiple QSs. Given the extremely high convergence of results caused by all three rules we applied, we expect that human coding could only improve reliability over the computer implementation of our random rules. As an independent test where human coding of all text units permits a broader test of reliability, we assess the reliability of natural sentence versus quasi-sentence coding using results of a series of experiments applying a CMP-based coding scheme to party manifestos for the European Elections conducted by Braun, Mikhaylov and Schmitt.[23] In this experimental design, expert text coders were randomly assigned to two groups. In a setup following Mikhaylov, Laver and Benoit,[24] both groups had to code the same excerpt of the 1999 British Liberal Democratic Party Euromanifesto using an online coding platform.[25] The first group (twenty-three participants) was asked first to unitize the document into QSs, and then to code these QSs using the coding scheme. The second group (twenty-nine participants) was asked simply to treat natural sentences as text units and code these using the coding scheme. Coders were undergraduate students from the University of Mannheim, followed coding instructions as outlined by Budge *et al.*,[26] and used a version of the CMP-based coding scheme modified to address issues specific to European Parliament elections.[27]

In order to assess the reliability and quality of the coding process and consequently of the data generation process, Braun, Mikhaylov and Schmitt calculate inter-coder agreement in each experimental group, with the aim of comparing the coding reliability of the two groups (which are defined by natural sentence versus quasi-sentence unitization).[28] Agreement was measured using Fleiss's kappa (κ).[29] The κ coefficient is by far the most widely used method of statistical analysis of agreement for categorical variables,[30] and generalizes directly to multiple coders rating the multiple items. The coefficient has a range from 0 (perfect disagreement) to 1 (perfect agreement), and takes into account the fact that some agreement may occur purely by chance.[31] While there are no universally accepted guidelines, it has been suggested that $\kappa > 0.75$ represents excellent agreement,

[23] Daniela Braun, Slava Mikhaylov and Hermann Schmitt, 'Computer-Assisted Human Coding: Experimental Tests of Reliability' (paper presented at 'Political Parties and Comparative Policy Agendas', an ESF Workshop on Political Parties and their Positions, and Policy Agendas, University of Manchester, 2010).

[24] Slava Mikhaylov, Michael Laver and Kenneth Benoit, 'Coder Reliability and Misclassification in the Human Coding of Party Manifestos', *Political Analysis*, 20 (2012), 78–91.

[25] The excerpt of the 1999 British Liberal Democrats Euromanifesto used in the experiment consists of 83 natural sentences. The Euromanifestos Project previously used this excerpt as the training document and declared it to consist of 112 QSs.

[26] Budge *et al.*, *Mapping Policy Preferences.*

[27] As part of the coding procedure, coders coded policy domains (the seven categories defined by the first digit of the CMP code) and coding categories sequentially.

[28] Braun, Mikhaylov and Schmitt, 'Computer-Assisted Human Coding'.

[29] Joseph L. Fleiss, 'Measuring Nominal Scale Agreement among Many Raters', *Psychological Bulletin*, 76 (1971), 378–83; Joseph L. Fleiss, Bruce A. Levin and Myunghee Cho Paik, *Statistical Methods for Rates and Proportions* (Hoboken, N.J.: Wiley, 2003).

[30] Chris Roberts, 'Modelling Patterns of Agreement for Nominal Scales', *Statistics in Medicine*, 27 (2008), 810–30, p. 811.

[31] The κ coefficient was proposed by Jacob Cohen ('A Coefficient of Agreement for Nominal Scales', *Educational and Psychological Measurement*, 20 (1960), 37–46) and extended to multiple raters by Fleiss ('Measuring Nominal Scale Agreement among Many Raters'). Hayes and Krippendorff compare Krippendorff's α and Fleiss's κ and state that they are very similar (Andrew F. Hayes and Klaus Krippendorff, 'Answering the Call for a Standard Reliability Measure for Coding Data', *Communication Methods and Measures*, 1 (2007), 77–89). We found that this also holds when applying the measures to the data collected by Braun, Mikhaylov and Schmitt ('Computer-Assisted Human Coding'). The κ coefficient is measured as $\kappa = (p_o - p_e)/(1 - p_e)$, where p_o is the overall proportion of observed agreement and p_e is the overall proportion of agreement expected by chance.

TABLE 3 *Inter-coder Reliability Results from the Experiment Comparing Natural and Quasi-Sentence Unitizations for the Euromanifesto Coding Scheme*

	Natural sentence		Quasi-sentence	
	κ	95% CI	κ	95% CI
Policy domain	0.397	(0.349–0.454)	0.384	(0.340–0.437)
Coding categories	0.315	(0.269–0.371)	0.313	(0.270–0.364)

Notes: Bootstrapped bias-corrected 95% confidence intervals from 1,000 replications. Bias correction gives better coverage probability for a possibly biased statistic, and produces the same results as the percentile method for an unbiased statistic. See Braun, Mikhaylov and Schmitt, 'Computer-Assisted Human Coding: Experimental Tests of Reliability'; and Anthony C. Davison and David V. Hinkley, *Bootstrap Methods and Their Application* (Cambridge: Cambridge University Press, 1997).

that $0.40 < \kappa < 0.75$ represents fair to good agreement, and that any $\kappa < 0.40$ indicates poor agreement.[32] Similar to our own results above, the QSs identified in the first stage of the Mannheim experiment were predominantly full natural sentences; furthermore, when one natural sentence consisted of more than one QS, these were typically coded into the same category. Just as with our own results, therefore, we would not expect much systematic difference in coding results between the two groups. The results relevant for our purpose are shown in Table 3.

The key result for our purposes is that the inter-coder reliabilities of the groups coding natural versus quasi-sentences is substantively indistinguishable. Coding reliability is quite poor in this experiment overall, possibly reflecting that the main objective of Braun, Mikhaylov and Schmitt was to test the reliability of a new coding scheme for the manifestos component of the European Election Study. Overall, however, the reliability coefficients reported in Table 3 are in line with the 0.3–0.4 from tests by coders applying the standard CMP coding scheme to pre-unitized QSs reported by Mikhaylov, Laver and Benoit.[33] In sum, there is no evidence in these results that coding natural sentences rather than QSs affects the reliability of coding. The huge gain in reliability from moving to an exogenous definition of text units does not appear to come at the expense of coding reliability.

CONCLUSIONS

It is fundamental to the systematic analysis of political text that we specify the basic unit of text analysis. The 'exogenous' specification of the natural sentence as the text unit results in perfectly reliable unitization, but potentially pays a price in coding validity if natural sentences contain more than one message. There is also a potential price in coding reliability if human coders are asked to pick one from several messages in a single natural sentence. Addressing these issues, we draw four primary conclusions.

First, in only a small minority of cases in the manifestos we examined are natural sentences divided into separate QSs. This means that in effect there is little possible difference between a scheme requiring humans to make painstaking and unreliable decisions on parsing natural sentences into smaller units, because most QSs are also natural sentence units.

Secondly, even when the QS unitization rules call for dividing a natural sentence into multiple text units, more than half of these subdivided natural sentences contained sub-units that all have the same code. This means that no information about alternative policy emphases is lost for these units

[32] J. Richard Landis and Gary G. Koch, 'The Measurement of Observer Agreement for Categorical Data', *Biometrics*, 33 (1977), 159–74; Fleiss, Levin and Paik, *Statistical Methods for Rates and Proportions*, p. 604.

[33] Mikhaylov, Laver and Benoit, 'Coder Reliability and Misclassification in the Human Coding of Party Manifestos'.

by considering only natural sentences. It also undermines the CMP rationale for using QSs, which is that one natural sentence may contain more than one distinct message, while it in effect results in double or triple counting of a subset of text units.

Thirdly, in our comparisons of the policy categories aggregated into percentages, including different left–right scales, we found no substantive differences between aggregations based on natural sentence versus quasi-sentence unitization. Our random procedure to assign a split natural sentence one of its constituent QS codes reproduced about 98 per cent of the variance in the aggregate measures based on QSs, including when subsamples were drawn to simulate the additional sampling variance that might come from having shorter manifestos. Because we think that human coders could improve on the random rules using expert judgement, furthermore, we expect our results to represent a worst-case scenario.

Finally, reporting the results from coder experiments where participants were asked to code either natural sentences or QSs, we found no evidence that inter-coder reliability differed between these two groups. The possible information loss from increasing the size of the text-coding unit from QSs to natural sentences does not on our evidence introduce additional unreliability.

The implication of these results for applying categorical coding schemes to political text is clear and simple. Natural sentences can be substituted for QSs to achieve a major gain in the reliability of text unitization without loss of validity. This implies that future text coding projects should dispense with endogenous text unitization by human experts as part of the coding process, and move to fully automated unitization based on natural sentence delimiters defined exogenously as part of the research design. Since our estimates suggest that substantive findings are unlikely to be affected by doing this, but reliability is likely to increase, the shift to natural sentence unitization could usefully be extended to the ongoing CMP and PA projects. Our analysis here implies that a substantial gain in reliability, efficiency and replicability can be achieved without sacrificing important substantive information in the texts under investigation.

APPENDIX TABLE A1 *Description of Manifestos Selected for Re-analysis*

Country	Party (CMP-code)	Election year	Language	CMP coder ID	Length in QSs (CMP)*	Party family (CMP)	Rile score (CMP)
Austria	People's Party (42520)	1971	German	104	213	CHR	25.3
Austria	Social Democratic Party (42320)	1979	German	104	541	SOC	−5.7
Australia	National Party (63810)	2001	English	201	173	AGR	18.5
Estonia	Fatherland Union (83710)	1999	Estonian	242	81	CON	22.2
Estonia	Fatherland Union (83710)	2003	Estonian	242	94	CON	24.5
Estonia	Res Publica (83611)	2003	Estonian	242	171	CON	−1.8
Britain	Conservative Party (51620)	2001	English	109	724	CON	14.9
Britain	Scottish National Party (51902)	2001	English	108	813	ETH	−13.0
Britain	Sinn Féin (51210)	2001	English	108	537	COM	−8.0
Iceland	Awakening of the Nation (15323)	1995	Icelandic	213	666	SOC	−11.4
Iceland	Independence Party (15620)	1978	Icelandic	212	100	CON	27.0
Ireland	Fine Gael (53520)	2007	English	280	2063	CHR	−9.7
New Zealand	ACT (64420)	1996	English	201	174	LIB	13.8
United States	Democratic Party (61320)	2000	English	204	1140	SOC	−3.6
United States	Democratic Party (61320)	2004	English	109	912	SOC	8.6

Party family abbreviations: AGR = Agrarian, CHR = Christian Democratic, COM = Communist, CON = Conservative, ETH = Ethnic-regionalist, LIB = Liberal, SOC = Social Democratic.

* The number of QSs recorded by the CMP in the coding of this document, e.g. the Austrian 1971 manifesto was determined by the CMP coder to contain 213 QSs.