# NATURAL SENTENCES AS VALID UNITS FOR CODED POLITICAL TEXTS[*]

Thomas Däubler
Trinity College Dublin

Kenneth Benoit
London School of Economics

Slava Mikhaylov
University College London

Michael Laver
New York University

8 April 2011

**Abstract**

Despite the recent focus on scaling policy positions by treating political text as quantitative data, huge investments in political science continue to use expert-coded content analysis, namely the 30-year Comparative Manifesto Project (CMP) of coded manifestos as well as the Comparative Policy Agendas Project (CAP). All text analysis methods require the identification of a fundamental unit of analysis. The fundamental unit of analysis in both CMP and CAP is the "quasi sentence", which is either a natural sentence, or a part of a sentence judged by the coder to have an independent component of meaning. The use of subjective judgment in identifying quasi-sentences, however, means that specification of the fundamental unit of data analysis is endogenous to the content of the text. In addition, it is known that the unitization of political texts into endogenous quasi sentences by expert coders generates unreliable specifications of the unit of analysis. The justification for using quasi-sentences is a supposed gain in associated validity of the codings. In this paper, we show that this justification is empirically questionable, since using quasi-sentences does not produce valuable additional information in characterizing substantive political content. Defining text units exogenously as natural language sub-units separated by one of a predefined list of punctuation marks, by contrast, generates perfectly reliable unitization, with no measurable cost in terms of the content validity of the resulting estimates.

**Key Words**: Political text analysis, policy positions, Comparative Manifestos Project, text unitization, content analysis, human coding.

[*] Prepared for presentation at the "Why and How of Party Manifestos in New and in Established Democracies" workshop of the 2011 ECPR Joint Sessions, 12-17 April, St. Gallen, Switzerland.

A rapidly growing area in political science has focused on perfecting techniques to treat political text as "data" to be analyzed, usually for the purposes of estimating latent traits such as left-right political policy positions (e.g. Laver et al. 2003; Slapin and Proksch 2008). More long-standing approaches have applied traditional content analysis tools to categorize sub-units of political text – such as sentences from manifestos – to measure the salience of political topics in party policy platforms. Prominent examples of these include the 30-year old Comparative Manifesto Project (Budge et al. 2001; Klingemann et al. 2006) as well as the Policy Agendas Project (Baumgartner et al. 2007). "Text as data" approaches convert text to purely quantitative information and use statistical tools to make inferences about characteristics of the political positions, sentiment, or topics represented in the text. Content analysis schemes employ humans to read textual sub-units and assign these to pre-defined categories. Both methods require the identification of a textual unit of analysis – a highly consequential, yet often unquestioned decision of research design – before the methods can be applied.

In this paper, we critically examine the dominant approach to unitizing political texts prior to human coding: the parsing of texts into *quasi-sentences*, defined as part or all of a natural sentence that express a distinct policy proposition. The use of the quasi-sentence rather than natural language units (such as sentences defined by punctuation) is motivated by the desire to capture all relevant political information, regardless of the stylistic decision to create long or short natural sentences. The correct identification of quasi-sentences by human coders, however, is highly unreliable. If we can demonstrate through comparing texts coded using both quasi- and natural sentences, that there is no appreciable difference in measured political content, then we would have a strong case for replacing human unitization schemes

with natural sentence text units that can be easily identified – and with perfect reliability – by computerized methods based on punctuation delimiters.

Our paper proceeds as follows. First, we discuss the main issues motivating the use of quasi-sentences and what this entails for reliability. Next, we re-examine and recode, using natural sentences, previously unitized and coded texts in several languages, and compare the aggregated coded results to see if this generates discernible differences in political content. Furthermore, we report results from a comparison of coding reliability from coding natural versus quasi-sentences. Our results provide overwhelming evidence suggesting that using natural language sub-units, in particular natural sentences, is always superior to unitization methods based on human judgment, and we issue recommendations accordingly.

## THE RATIONALE FOR ENDOGENOUSLY DEFINED TEXT UNITS

Expert or "hand" coded political text is certainly not alone in facing the issue of how to define the unit of analysis for textual research. *Statistical scaling methods*, in which there have been numerous recent advances (Laver et al. 2003; Slapin and Proksch 2008), typically make the linguistic "bag of words" assumption and consider the atomic word as the unit of textual analysis. All substantive decisions about texts are made as part of the research design, not the coding process. *Dictionary-based methods* apply a predefined 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 in scaling methods, the goal is typically fully automated machine coding, with all substantive decisions made as part of the research design, and most especially during dictionary development, 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 examples can be found in Google Translate, or the Watson system recently and successfully developed

by IBM to understand and then answer the complex questions that form part of the Jeopardy TV quiz program. The unit of text analysis is now much more than a word or n-gram and may be endogenous. Thus far this research program has been the preserve of computer scientists and computational linguists, as opposed to political scientists.

In expert text coding, the method of research projects such as the Comparative Manifesto Project and the Comparative Policy Agendas Project, the objective is basically non-automated natural language processing. Crudely speaking, expert coding can be seen as the employment of skilled humans to engage in complex pattern recognition tasks that we cannot yet program 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 be endogenous to the meaning of the text. Because only human judgment can (yet) be trusted to provide valid results, such traditional methods of content analysis inherently involve subjective judgment by humans reading, parsing, and coding the text. This method of introducing human judgment as part of the research process rather than just the research design, however, introduces unique concerns about reliability that are not issue with the other methods of textual analysis we have identified.

## RELIABILITY TRADEOFFS WITH SUBJECTIVE TEXT UNITIZATION

Expert text coding involves two basic data-generating steps (Krippendorff 2004, 219). First, the text is *unitized* by dividing it into smaller units relevant to the research question, such as quasi-sentences, although it would also be possible to identify less subjective units such as words, sentences, paragraphs, or pages. Unitization can be defined *exogenously* to the research process, such as identifying units according to syntactical rules or of fixed word length, that involve no human judgment during the application of content analysis.

Alternatively, they may be defined *endogenously* to the process, involving human judgment to determine where one unit of content ends and another begins as part of the content analysis itself. 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.

The core issue when designing schemes for both unitization and coding of text units for an expert coding project is the classic trade-off between reliability and validity. In expert text coding a research procedure is reliable when "the reading of textual data as well as of the research results is replicable elsewhere, that researchers demonstrably agree on what they are talking about" (Krippendorff 2004, 211). Validity, at its simplest, means that the results of the content analysis will reflect the true content of the text in a meaningful way. If expert coders must apply a subjective scheme to classify the content of units of a text, then they almost surely have chosen this laborious route over machine coding because they feel the results are more valid – that humans can currently extract more valid meaning from complex texts than can machines. Furthermore, expert codings are likely most valid when the unit of text analysis is endogenous, since it is unlikely that readers for whom the texts are written pay close attention to punctuation marks when they read a text for meaning.

The two most widely used coding schemes in political science – the Comparative Manifesto Project (Budge et al. 2001) and the Comparative Policy Agendas Project (Baumgartner et al. 2007) – both specify the unit of textual analysis as an endogenous text fragment known as the *quasi-sentence*: "an argument which is the verbal expression of one political idea or issue" (Volkens 2001, 96). This approach to unitizing is often referred to as thematic unitizing (Krippendorff 2004). The explicit motivation for using quasi-sentences 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.

**[FIGURE 1 ABOUT HERE]**

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 having strong families and communities." The CMP coder of this document identified two quasi-sentences, 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-sentence 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 quasi-sentences by a coder from the Comparative Manifesto Project. Quasi sentences are demarcated by the pencil marks in the text, indicating that ten natural sentences have been divided into 23 quasi-sentences, some as short as a single word. It should be clear to any reader considering this example that it is 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 for identifying an alternative set of word strings qualifying as independent quasi-sentences from this short manifesto fragment.[1]

**[FIGURE 2 ABOUT HERE]**

Carefully training expert coders to follow well-defined instructions may mitigate the problems of unitization unreliability, but experience shows that it can hardly alleviate them.

[1] To return to the Australian 2001 National Party example cited earlier, we also observe the following the 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 quasi-sentence and coded it as 501 (Environmental Protection: Positive), even though it could plausibly have been seen as comprising two quasi-sentences, divided by the "and", with the first coded to 410 (Productivity: Positive) and the second to 501.

In data we obtained from the CMP from their own coder training experiments, wherein expert coders were asked to unitize and code a training document, the results from 67 trained coders showed huge variability in the number of quasi-sentences they identified in the text. Figure 2 depicts (as a kernel density estimate) the distribution of the total quasi-sentences identified as independent text units by 67 trained expert coders. In the CMP's master coding, applying the authoritative version of the quasi-sentence unitization scheme, the document contains a "true" number of 163 quasi-sentences. The expert coders, however, identified a total number of quasi-sentences ranging from about 120 to 220, with a standard deviation of 19.[2] When even well-trained human expert coders specify units of analysis endogenously, and this is precisely what CMP coders do when they parse a text into "quasi-sentences", the results are extremely unreliable. Some expert coders find many more quasi-sentences in precisely the same text than do others, while others find fewer. Unlike in our example here – and arguably, not even in this one – there is no "gold standard" for assessing which expert has made the correct unitization and which has made the wrong choices. Because human-coded content analysis schemes almost always combine results from different coders, furthermore, any systematic differences in subjective judgment about what constitutes a proper endogenously identified text unit are likely to be correlated with particular texts, introducing possible bias as well as additional uncertainty.

An ideal solution to the need to balance reliability and validity would be to develop a reliable, automated method for identifying independent quasi-sentences. The hard problems of natural language processing for complex political texts, however, mean that no automated unitization of texts into quasi-sentences is currently feasible – at least, none in which we

[2] In the test results, coders with especially bad first round results had these corrected, and were asked to repeat the experiment. In Figure 2, 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.

would confidently declare is valid in making the information-rich thematic distinctions motivating the use of endogenous text units. An alternative is a more efficient and reliable 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 (Krippendorff 2004, 104). Among the choices of syntactically delimited units, *natural sentences* are closest to the thematically defined quasi-sentences used by 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 paper, concerns whether specifying the unit of analysis as exogenously specified natural sentences, rather than endogenously specified quasi-sentences, significantly affects inferences about the substantive content of the types of text we wish to investigate. 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 embracing perfectly reliable natural sentences as the unit of textual analysis for expert coding is a dominant methodological strategy.

## DATA AND METHODS

Our comparison of validity of expert-coded text analysis based on exogenous versus endogenous text units comes from a reanalysis of manifestos originally unitized and coded according to the Comparative Manifesto Project and CMP-inspired schemes. 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 quasi-sentences, 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.[3] 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.

Such a test would be expensive and time-consuming to design, so we have settled for two other tests involving larger numbers of manifestos. The first major set of tests involves returning to manifestos that have been previously unitized into quasi-sentences and coded by trained CMP coders, and indeed form the data reported in the CMP dataset. This set of 13 documents consists of printed manifestos with unitization marks and marginal codes of the sort depicted in Figure 1. Our approach proceeds in two steps. First, we recorded all quasi-sentence codes indicated on the margin of the documents, plus the information if these are identical to natural sentences or which of the quasi-sentences are components of the same natural sentence. In essence, this yields a dataset where the unit of the analysis is the natural sentence and component quasi-sentences (one or several) are sub-units.[4] 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 only a single quasi-sentence.* In this case, the policy code for the natural sentence is assigned that of the quasi sentence.
2) *A natural sentence contains more than one quasi-sentence but all have the same policy codes*. In this case, the natural sentence receives the same code.

[3] 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.

[4] 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.

3) *A natural sentence contains more than one quasi-sentence, and these have different policy codes*. In this case, if a human were coding the natural sentence and faced with this choice, she or he would probably decide which of the possible competing policy codes best represented the natural sentence unit, and choose that code. Our procedure in this case uses even less information – and is hence more conservative in the sense that any real expert could almost certainly produce better results. Our procedure applied three different rules for choosing among competing quasi-sentence codes to assign to the natural sentence:

   *a) First*: Assign the natural sentence the code of the first component quasi-sentence.

   *b) Last*: Assign the natural sentence the code of the last component quasi-sentence.

   *c) Random*: Assign the natural sentence the code of a randomly chosen component quasi-sentence.

In a remarkable (and uncharacteristic) display of compassion for graduate student assistants, the authors themselves carefully applied this method to a total of 13 manifestos (so far) from a variety of political contexts and written in different languages. The sample includes five English texts, consisting of one manifesto from Australia, one from New Zealand, two from the UK and two from the US; three Estonian manifestos (in Estonian); two German-language manifestos from Austria; and two manifestos from Iceland (in Icelandic).

Our analysis has two main aims. First, we use the data to describe how frequently the three types of relationships between natural and quasi-sentence units occur. Second, by comparing the aggregated political content in the form of comparisons of the aggregate proportion of each policy category or more inclusive policy indexes such as a left-right or an environmentalism scale, we assess if our substantial conclusions about document content

change when we shift to natural sentences (comparing the *first*, *last* and *random* rules for assigning a code to the natural sentence level in the third type of situation).

As previously mentioned, of course, humans faced with a natural sentence that clearly does contain more than one policy statement may face a tough choice in deciding which code to assign it, if only one code may be assigned. In such a case, it is possible that coder reliability – a separate issue from unitization reliability – may be adversely affected. 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 painstaking revisiting and recording of the text units from the 13 manifestos provided a dataset of a total of 4,859 natural sentences, in which were contained 5,660 quasi-sentences. These are described in Table 1.

**[TABLE 1 ABOUT HERE]**

The clearest result to emerge from our analysis is that the splitting of natural sentences into more than one quasi-sentence by CMP coders occurs quite infrequently: In almost nine out of ten cases (88.5%), natural sentences contained only a single quasi-sentence, meaning that all this fuss pertains to fewer than just 12% of all text units. The remaining natural sentences include mostly two (8.4% of all natural sentences) or three quasi-sentences (2.0% of all natural sentences). Natural sentences with four or more quasi-sentences are very rare, making up just 1% of our sample.

The second strong result from our analysis is that when natural sentences are split into component quasi-sentences, these components are not necessarily coded differently. In fact, in the category of natural sentences with two quasi-sentences, less than half (43.2%) of the natural sentences have different component codes, rising to just over half (53.5%) for natural sentences split into three quasi-sentence units. More of those split into four or more quasi-sentences were different, although overall, as previously mentioned, these represent just a tiny fraction of all of natural sentences. If we consider all natural sentences as the total, the overall share of cases with varying component codes is just 5.4%. In other words, before any additional comparison, we expect results that are at the very least 95% identical, because there is a 95% similarity between the two unitizations.

What is more, this share of 5.4% refers to differences judged on the basis of the 56-category-CMP-scheme.[5] Secondary analyses typically work at a more aggregate level. By a country mile, the most popular application of the CMP data is the use of the left-right index "Rile", a scheme that considers nearly a fourth (13 of 56) of the CMP categories as "right", another fourth as "left", and the rest as neither. Since it is plausible that natural sentences split among differently coded quasi-sentences might still contribute in the same way to the Rile index, by virtue of still belonging to the same left, neutral or right "Rile" category, we also analyzed the splits according to this highly simplified three-category scheme. The fifth column of Table 2 shows that the share of component quasi-sentences in disagreement further drops if we base our judgment on this three-fold classification. Among the natural sentences with two component quasi-sentences, less than a third (31.9%) of cases feature within-

[5] 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 5,660 (0.12%) quasi-sentences. In addition, 24 quasi-sentence codes (0.42%) could not be identified from the documents since they were not legible.

sentence differences in terms of a left-right-neutral classification. The total share of natural sentences with component codes that differ in terms of this orientation is only 3.9%.

The results from Table 1 are based on the pooled natural sentences data. This might conceal differences across manifestos, also because longer documents will contribute a higher share of cases to the data. Table 2 therefore presents the results of the same analysis, but aggregated at the manifesto level. As stated above, we have coded 13 documents so far, which form the units of analysis in this table.

**[TABLE 2 ABOUT HERE]**

Table 2 illustrates two main points. First, if we consider a typical manifesto as represented by the median or mean, the substantial conclusions are the same as for Table 1. Respective figures are slightly higher than in Table 1, but (based on the mean) only 14.2% of natural sentences in a document include more than one quasi-sentence, only 7.3% of natural sentences contain different component codes and only 5.5% are cases where component codes differ in orientation (left, right, neutral). The second point is that there is considerable variation in these measures across documents. For instance, the share of natural sentences that are split into quasi-sentences ranges from 1% (Independence Party, Iceland, 1978) to 46% (Scottish National Party, UK, 2001). To which extent these differences are driven by variation in the nature of the political text or simply by inter-coder variation in propensity to split natural sentences cannot be answered with these data (at least not at the moment).

In a purely descriptive sense, our analysis comparing natural to quasi-sentence units has shown that even prior to our comparison of aggregate political content, we would expect similarities of 95 and 96 percent between coding based on perfectly reliable exogenously defined text units – natural sentences – and unreliable, labor-intensive endogenously defined text units – quasi-sentences – because in practice these units of analysis are exactly the same in 19 out of 20 cases.

*Comparing Aggregate Results*

Individual sentence codings are not only of little substantive interest to end users of political content analysis datasets, they are not even reported. Instead, the CMP dataset contains only the percentages of each policy category – the "per" codes – as well as the total number of quasi-sentences recorded in the manifesto. Our comparison in this section is therefore to compare aggregate category percentages from each manifesto when these are reconstructed from quasi- and natural sentences.[6] This involved applying our three coding rules – choose the first quasi-sentence code, the last, or one at random – to code the natural sentence. As we have emphasized already, this affects just 5.4% of all natural sentences.

**[FIGURE 3 ABOUT HERE]**

Figure 3 shows the comparison of each policy category's percentage share, in a scatterplot matrix comparing the three rules to the quasi-sentence-based results. Each point represents a policy percentage from one manifesto, and the dashed line shows the 45-degree axis of perfect agreement. To reduce some of the skew created by low-frequency policy categories, we have logged both axes (although this makes no difference to the results). The squares above the diagonal report Pearson's correlation coefficient, ranging from 0.98 to 0.99 – 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 quasi-sentence policy category percentages on the logged natural sentence policy category percentages. The results confirm those from the scatterplots, indicating that 98% of the variance in the original quasi-sentence coding is explained by the natural sentence codings, regardless which rule is applied. An *F*-test whether the estimated slope coefficient

[6] The emphasis here is on "reconstructed": we did not ensure that every category percentage from the quasi-sentences 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.

differs from the 1.0 value of perfect identity, furthermore, cannot reject this null hypothesis. All told, this is strong, incontrovertible evidence that the natural sentence and quasi-sentence codings yield the same aggregate results. Even with the random assignment rules – over which smart humans reading the natural sentence could presumably improve – our similarity has risen from the baseline of being 95.4% identical to at least 98% identical.

**[TABLE 3 ABOUT HERE]**

Of course, individual policy categories are seldom used directly by applied researchers. Instead, this honor falls to the left-right "Rile" index that includes 26 of the 56 total categories. In Figure 4, we show the aggregate results on the Rile index for our 13 manifestos, indicating an extremely high degree of agreement. Because Figure 4 only has one data point for each of the manifestos for which we painfully, eye-wateringly recoded the text units, we re-sampled natural sentences from each manifesto and plotted these in Figure 5. In this figure, we drew 100 natural sentences from each manifesto 100 times each, to plot a total of 13 x 100 = 1,300 points representing hypothetical, shorter manifestos drawn from the 13 manifestos in our sample. The top panel of Figure 5 shows the CMP's additive, original Rile scale, while the bottom panel depicts the aggregate logit Rile scale proposed by Lowe et al (2011),[7] a scale that has been argued has better properties than the CMP's relative difference scale. In both cases, almost perfect correspondence is observed, even given the variation to be expected in each case from the sampling procedure.

**[FIGURES 4 and 5 ABOUT HERE]**

The left-right index uses a large number of categories (26 of 56) and previous comparisons (e.g. Lowe et al. 2011) have shown that it is fairly robust to different computations. To test the aggregate differences on policy category with typically smaller

[7] 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.

frequencies of coded policy statements, we also applied the re-sampling procedure to test differences in aggregate environmentalism scores, using the logit environmentalism scale from Lowe et al (2011).[8] The results, with only minor exceptions due to sampling variability, provide strong evidence of a near-perfect correspondence in results.

**[FIGURE 6 ABOUT HERE]**

## RESULTS FROM THE CODING RELIABILITY EXPERIMENT

Our results from above suggest that using natural sentences as units of analysis does not affect the validity of the classification of these units. Our test applied a random assignment procedure to code natural sentences, when these sentences were split into differently coded quasi-sentences. Of course, a computer applying deterministic or random rules to do this will suffer no qualms of indecisiveness or display no favoritism toward particular policy categories or domains. It is conceivable, however, that a human faced with a natural sentence clearly containing two separate, and distinct, policy statements will not use a consistent rule in coding a larger, natural sentence text unit. While eliminating the unreliability of subjective unitization, it remains to be tested whether we are not also increasing the unreliability of coding by forcing coders to make a Sophie's Choice on text units that could and perhaps should be considered to express more than one competing policy statement.

To test whether this is the case, we report here the results of a series of experiments conducted by Braun et al (2010) to apply the CMP coding scheme applied to European manifestos. In this experimental design, expert text coders were randomly assigned to two groups. In a setup similar to Mikhaylov et al (2010), both groups had to code the same

[8] This is computed as the Rile scale where the pro-environmental "R" categories is the sum of 501 Environmental Protection: Positive and 416 Anti-Growth Economy: Positive, and "L" is 410 Productivity: Positive. We excluded three manifestos (64420 and 83710 from 1999 and 2003) because these had no or only a single environment-coded text unit.

excerpt of the 1999 British Liberal Democratic Party Euromanifesto[9] using online coding platform. The first group (23 participants) was asked, first, to unitize the document into quasi-sentences, and to these quasi-sentences in the second stage. The second group (29 participants) was assigned to code text that was pre-unitized using natural sentences. This reflects a more complex decision making process of coders using thematic unitizing, where they first have to identify quasi-sentences and then code them according to the coding scheme. While coders using natural sentence unitizing use syntactic cues and proceed to coding almost immediately moving from one natural sentence to the next. Both groups were comprised of the undergraduate students from the University of Mannheim, from different academic fields and at different stages of their studies. Participants followed coding instructions presented in Budge et al (2001), and used the CMP-based coding scheme that was modified to address some issues that are specific to the European Parliament elections.

In order to assess reliability and quality of the coding process and consequently of data generation process, Braun et al (2010) calculate inter-coder agreement in each experimental groups. Estimating inter-coder agreement for each group individually allows comparing coding reliability in both experimental groups on a common scale. Furthermore, it mitigates some of the problems of additional uncertainty of thematic unitizing faced by the group using quasi-sentences. The uncertainity arises from the fact that the group that used thematic unitizing was prone to the same unitizing uncertainty shown in Figure 2 above. However, just like with the results presented in previous section, quasi-sentences identified in the first stage of the experiment were predominantly the same as natural sentences, and where one natural sentence consisted of more than one quasi-sentence these were almost always

[9] 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 of112 quasi-sentences.

coded into the same categories. Thus, similarly to our results above, we would not expect much systematic difference in coding reliability between two groups.

Braun et al (2010) use Fleiss's kappa (Fleiss 1971; Fleiss et al. 2003) to measure inter-coder agreement. Kappa coefficient has a range from zero (perfect disagreement) to one (perfect agreement), and takes into account the fact that some agreement may occur purely by chance. As part of the coding procedure, coders were coding policy domains (the seven categories defined by the first digit of the CMP code) and coding categories sequentially. The results of the experiment relevant for our purpose are presented in Table 4.

**[TABLE 4 ABOUT HERE]**

Setting aside the fact that the inter-coder reliability for both groups is abysmally poor – well below the 0.6 conservatively considered a standard of reliability for social sciences, or the 0.8 considered the point below which you would sue your radiologist – we see clearly that the inter-coder reliabilities of the two groups are both statistically and substantively indistinguishable. Coding reliability is admittedly terrible in this experiment, but not far below the 0.45-0.55 reported from tests by trained CMP coders in Mikhaylov et al (2010). Comparing the groups, we find evidence that being forced to code natural sentences instead of more focused quasi-sentences does not adversely affect the reliability of coding. The massive gain in reliability from the move to an exogenous definition of text units, in other words, does not come at the expense of coding reliability.

## CONCLUSIONS

Based on these tests, we draw three primary conclusions. First, in only a small minority of cases in the manifestos we examine are natural sentences divided into separate quasi-sentences. 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, simply because most "quasi" sentences are also natural sentence units.

Second, even when the quasi-sentence unitization rules call for dividing a natural sentence into multiple text units, more than a half of these subdivided natural sentences (53%) contained sub-units with all the same code. This means that no information about alternative policy emphases can be lost for these units by considering only natural sentences.

Third, in our comparisons of the policy categories aggregated into percentages, including indexes of left-right and environmentalism, we found no substantive differences between aggregations from natural versus quasi-sentence text units. Our random procedure to assign a split natural sentence one its constituent quasi-sentence codes reproduced about 98% of the variance in the aggregate measures based on quasi-sentences, with similar results from the index measures, 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 judgment, 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 quasi- or natural sentences, we found no evidence that reliability between these two groups was different. The possible information loss from increasing the size of the text coding unit from quasi- to natural sentences, in other words, does not appear to introduce additional unreliability.

The implication for applying categorical coding schemes to political text is a clear and simple lesson: natural sentences can be substituted for quasi-sentences without any loss of validity, even when no modification of the coding scheme itself is considered. Moving forward, it also implies that future coding schemes can dispense with endogenous unitization methods grounded in human-based decision, and move to fully automated methods based on

natural sentence delimiters. Our analysis here proves that this massive gain in reliability, efficiency, and replicability can be gained without sacrificing any important substantive political information.

## REFERENCES


Baumgartner, Frank R., Christoffer Green-Pedersen, and Bryan D. Jones. 2007. *Comparative Studies of Policy Agendas*. London: Routledge.

Braun, Daniela, Slava Mikhaylov, and Hermann Schmitt. 2010. "Human Coding of Party Programs: Experimental assessment of unitizing reliability." In *Political Parties and Comparative Policy Agendas: an ESF Workshop on Political Parties and their Positions, and Policy Agendas*. University of Manchester, May 20-21.

Budge, Ian, Hans-Dieter Klingemann, Andrea Volkens, Judith Bara, Eric Tannenbaum, Richard Fording, Derek Hearl, Hee Min Kim, Michael McDonald, and Silvia Mendes. 2001. *Mapping policy preferences : estimates for parties, electors, and governments, 1945-1998*. Oxford ; New York: Oxford University Press.

Fleiss, J.L. 1971. "Measuring nominal scale agreement among many raters." *Psychological Bulletin* 76 (5):378-383.

Fleiss, Joseph L., Bruce A. Levin, and Myunghee Cho Paik. 2003. *Statistical methods for rates and proportions*. 3rd ed. Hoboken, N.J.: J. Wiley.

Klingemann, Hans-Dieter, Andrea Volkens, Judith Bara, Ian Budge, and Michael McDonald. 2006. *Mapping policy preferences II : estimates for parties, electors, and governments in Eastern Europe, European Union, and OECD 1990-2003*. Oxford: Oxford University Press.

Krippendorff, Klaus. 2004. *Content analysis : an introduction to its methodology*. 2nd ed. Thousand Oaks, Calif.: Sage.

Laver, Michael, Kenneth Benoit, and John Garry. 2003. "Extracting Policy Positions from Texts Using Words as Data." *American Political Science Review* 97 (2):311-331.

Lowe, Will, Kenneth Benoit, Slava Mikhaylov, and Michael Laver. 2011. "Scaling Policy Preferences From Coded Political Texts." *Legislative Studies Quarterly* 36 (1):123-155.

Mikhaylov, Slava, Michael Laver, and Kenneth Benoit. 2010. "Coder Reliability and Misclassification in Comparative Manifesto Project Codings." In *Previously presented at the 66th MPSA Annual National Conference*. Palmer House Hilton Hotel and Towers.

Slapin, Jonathan, and Sven-Oliver Proksch. 2008. "A Scaling Model for Estimating Time-Series Party Positions from Texts." *American Journal of Political Science* 52 (3):705-722.

Volkens, Andrea. 2001. "Quantifying the election programmes: Coding procedures and controls." In *Mapping Policy Preferences: Parties, Electors and Governments: 1945-1998: Estimates for Parties, Electors and Governments 1945-1998*, ed. I. Budge, H.-D. Klingemann, A. Volkens, J. Bara, E. Tannenbaum, R. Fording, D. Hearl, H. M. Kim, M. McDonald and S. Mendes. Oxford; New York: Oxford University Press.

| Number of Quasi-sentences contained within Natural Sentences | *N* Natural Sentences | % Natural Sentences | % of Natural Sentences with Different CMP Codes | % of Natural Sentences with Different Right-left Codes (left, neutral, right) |
|---|---|---|---|---|
| One | 4,302 | 88.5 | (0) | (0) |
| Two | 407 | 8.4 | 43.2 | 31.9 |
| Three | 99 | 2.0 | 53.5 | 36.4 |
| Four | 26 | 0.5 | 73.1 | 53.9 |
| More than four | 25 | 0.5 | 52.0 | 36.0 |
| Total | 4,859 | 100 | 5.4 | 3.9 |

Table 1. *Pattern of Natural Sentences versus Quasi-Sentences from 13 election manifestos.*

| | Min | 1st quartile | Median | Mean | 3rd quartile | Max | Total manifestos |
|---|---|---|---|---|---|---|---|
| Split into several quasi-sentences | 1.0 | 2.3 | 14.1 | 14.2 | 21.3 | 46.0 | 13 |
| Split into several quasi-sentences that have… | | | | | | | |
| *differing component codes* | 1.0 | 1.7 | 6.4 | 7.3 | 11.5 | 21.5 | 13 |
| *component codes differing in orientation (left, right, neutral)* | 1.0 | 1.3 | 3.9 | 5.5 | 8.2 | 15.0 | 13 |

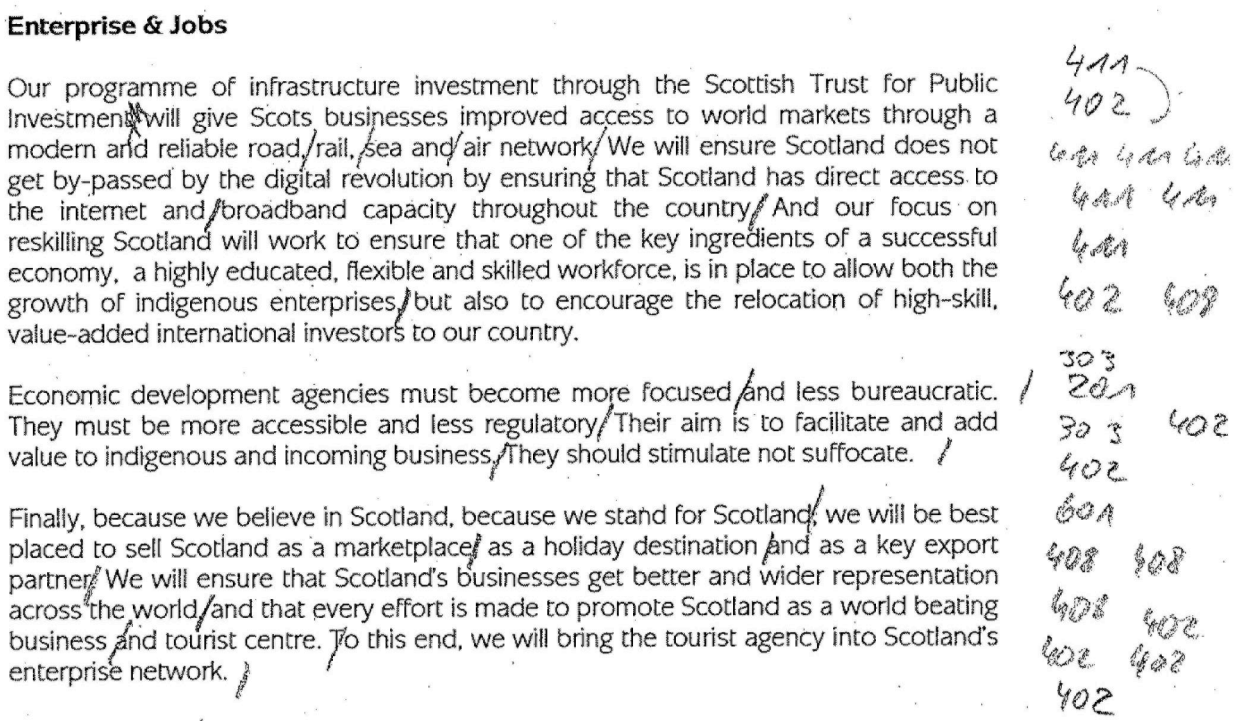
Table 2. *Characteristics of natural sentences at the manifesto-level (share in %)*

| Dependent variable: log(Quasi-sentence per) | | | |
|---|---|---|---|
| | (1) Random QS Code | (2) First QS Code | (3) Last QS Code |
| log(Natural sentence per) | **0.993** | **0.989** | **0.989** |
| | (0.008) | (0.008) | (0.008) |
| *N* | 376 | 380 | 377 |
| $R^2$ | **0.98** | **0.98** | **0.98** |
| *p*-value for *F*-test that β=1.0 | 0.37 | 0.16 | 0.41 |

Table 3. *Regression of (log) Quasi-sentence-based % categories by manifesto on (log) natural sentence-based estimates using three rules*. 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.

| | Natural Sentence | | Quasi-sentence | |
|---|---|---|---|---|
| | Kappa | 95% CI | Kappa | 95% CI |
| Policy domain | 0.397 | (0.343 - 0.457) | 0.384 | (0.335 - 0.441) |
| Coding categories | 0.315 | (0.260 - 0.365) | 0.313 | (0.269 - 0.360) |

Table 4. *Inter-coder reliability results from in Braun et al (2010) experiment of the Euromanifesto coding scheme for natural and quasi-sentence unitizations.* Bootstrapped 95% confidence intervals from 500 replications.



**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 internet 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.

Figure 1. *Section of SNP 2001 manifesto parsed into quasi sentences by CMP coder*

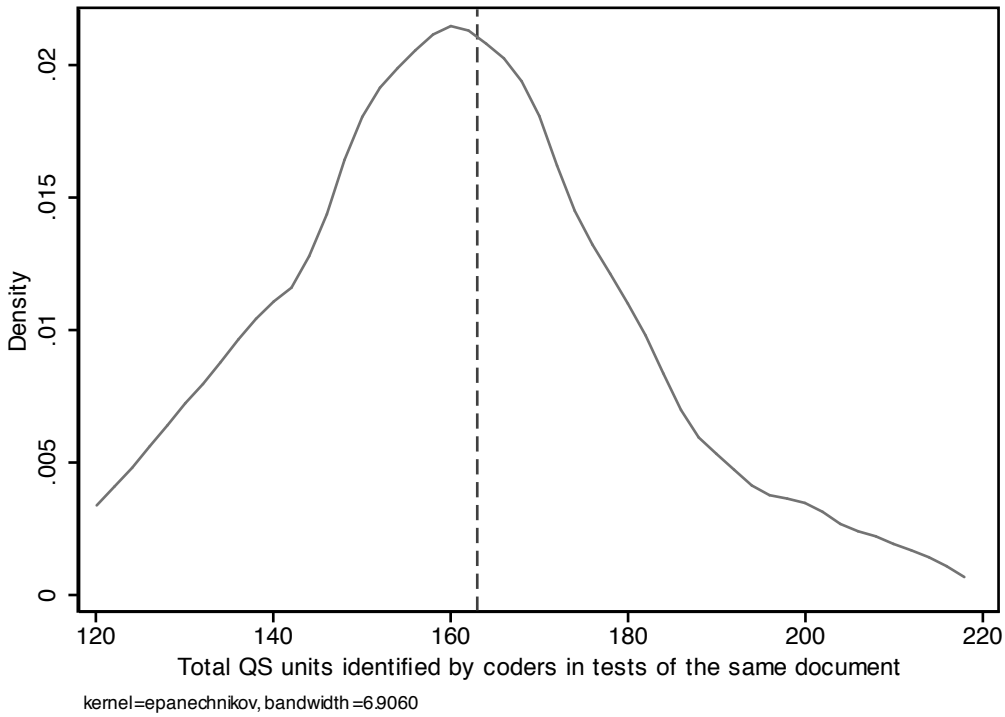



Figure 2. *Density plot of the total number of quasi-sentences identified in a CMP training text by 67 trained coders.* Source: Andrea Volkens.

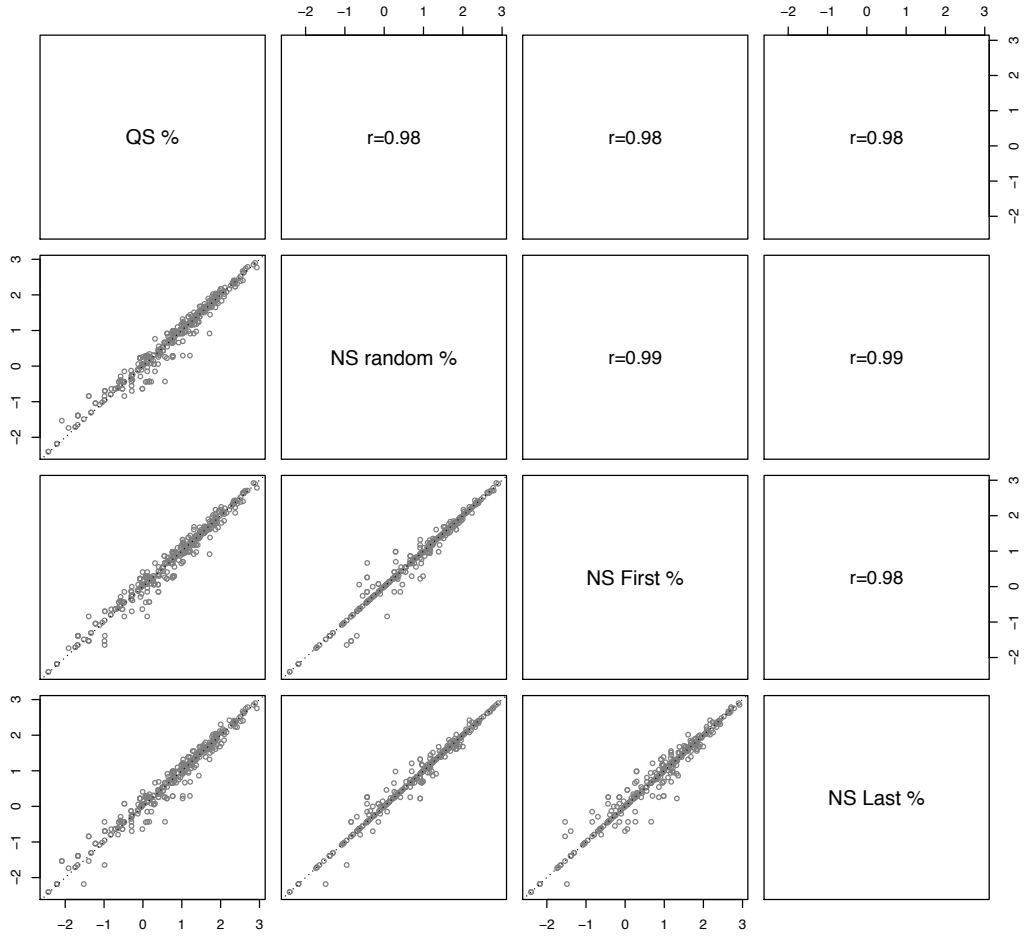



Figure 3. *Comparing quasi-sentence aggregate category percentages to natural sentence recodings.* Three rules are compared: randomly assign the code based on constituent quasi-sentences; take the first QS code for the natural sentence; and take the last QS code for the natural sentences. Total manifestos analyzed: 13.

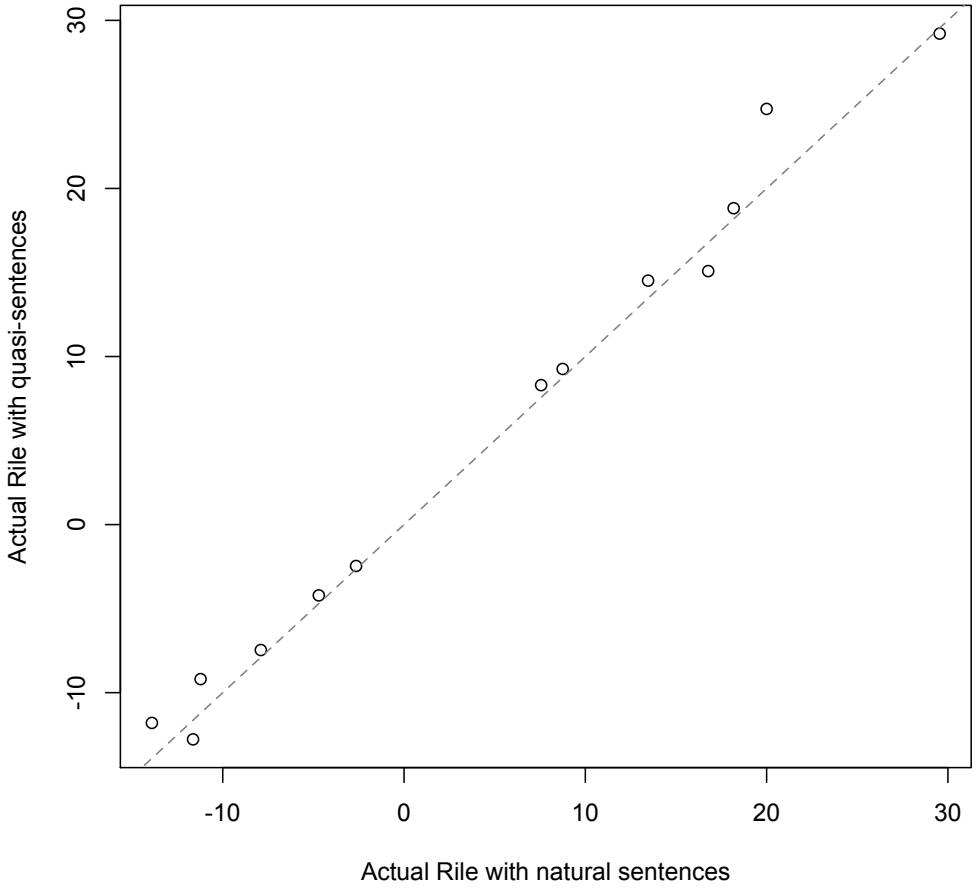



Figure 4. *Actual Rile Values aggregated for each manifesto.* Based on random assignment, which we have chosen because human coders could almost certainly do better than this rule.

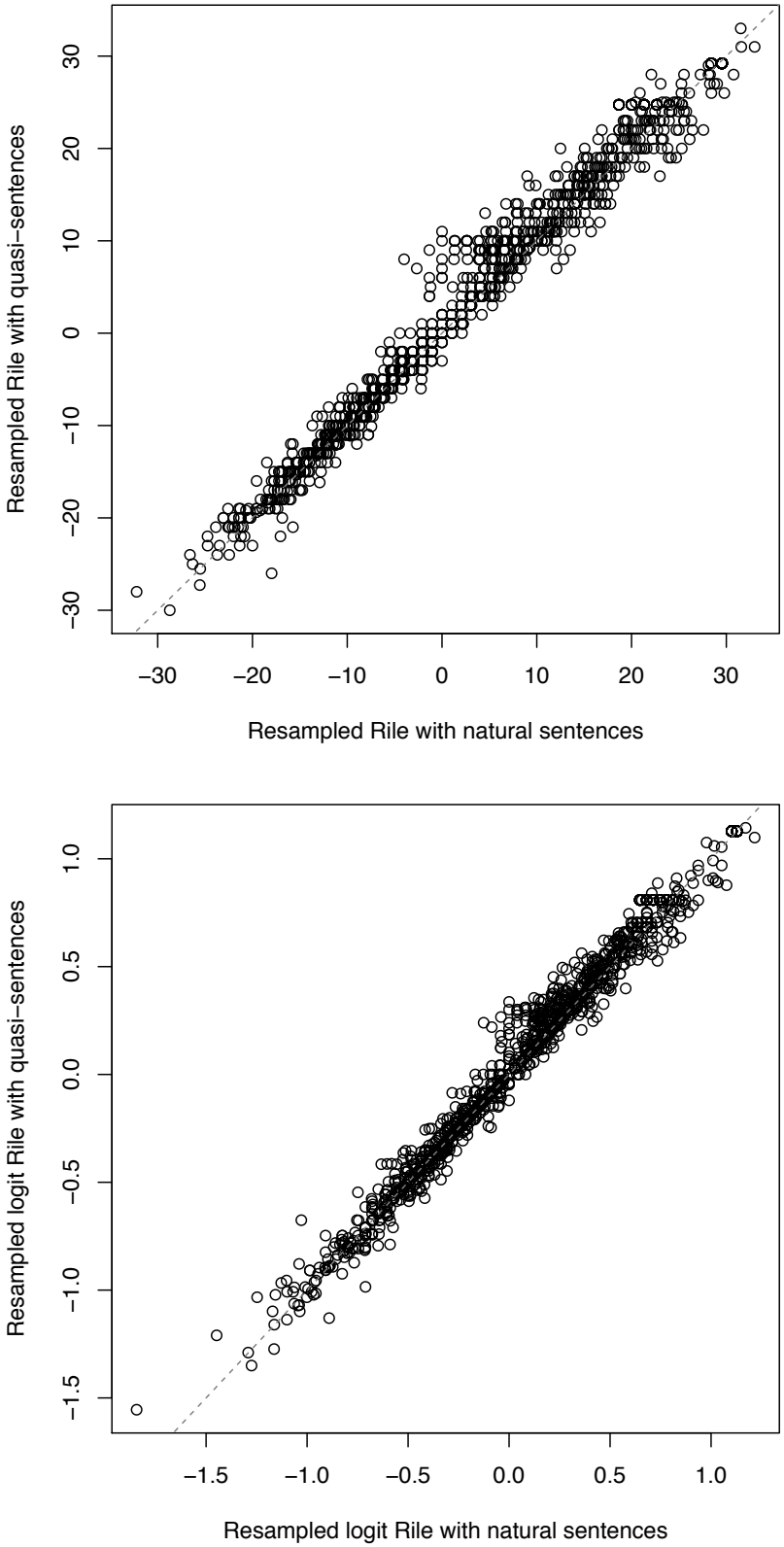



Figure 5. *Resampled Rile Values.* Based on random assignment, we took 100 random draws of 100 natural sentences each from each manifesto, and plotted the overall distribution of scores. The bottom plot uses the log Rile from Lowe et al (2011).

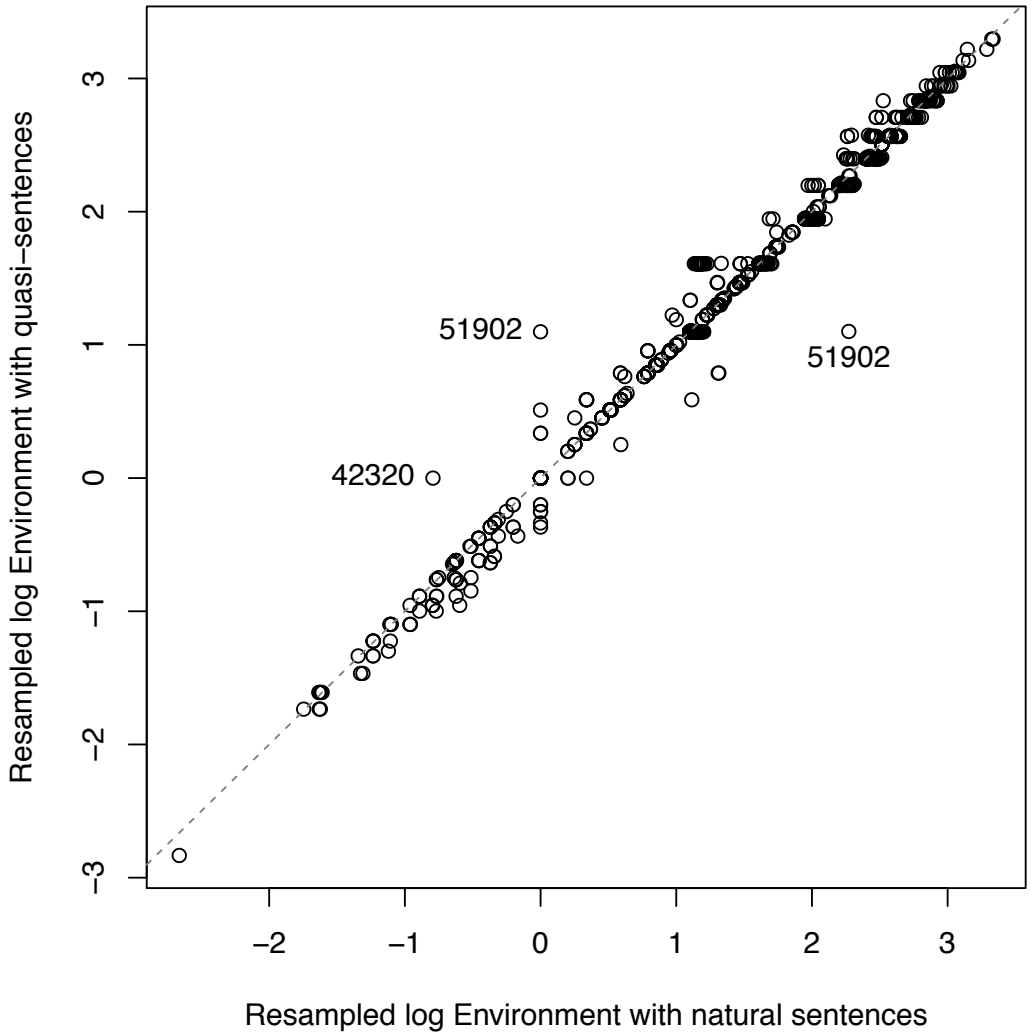



Figure 6. *Resampled Log Environment Scores.* Random resampling as per Figure 5, using the logit scale of environmental policy from Lowe et al (2011). The labelled manifestos are those with very low environmental content. We excluded 64420 and 83710 because these had extremely low (or none in the Estonian case) environmental quasi-sentences.