# CODING INSTRUCTIONS AND DATASET DESCRIPTION FROM MANIFESTO TEXT UNITIZATION STUDY

Kenneth Benoit, Thomas Daubler, Michael Laver, and Slava Mikhaylov
6 April 2011

## SUMMARY

These instructions describe the manner in which coded manifesto texts were recorded for the text unitization paper by Benoit, Daubler, Laver, and Mikhaylov.

Each text to be coded consists of a scanned manifesto text that has been coded by a trained CMP coder and includes the (hand-marked) codings. The coder markings indicate a) unitization of the text into quasi-sentences,[1] and b) assignment of one of the 56 CMP codes to each quasi-sentence, marked in the margins.

## DATASET

The dataset consists of *coded quasi-sentence (QS) units*. The codes to be recorded for each manifesto are:

- `party` – The CMP's 5-digit code for the party (see *MPP2* 2006).
- `year` – The 4-digit year (e.g. "1990").
- `language` – ISO abbreviation of the language of the manifesto text (e.g. "EN", "DE", "ES" for English, German, and Spanish respectively). This will most of the time (but not necessarily!) be the same as the first part of the filename.
- `pagesno` – the serial page number of the (pdf or printed) manifesto on which the unit is located. At least when coding from pdf document, use the pdf document page (and not the original page number of the manifesto).
- `QSpagesno`– The serial number of the QS from the current page. This will start at 1 with each new page (see below for instructions about quasi-sentences that span two pages).
- `NSpagesno` – The serially numbered natural sentence from the current page in which the quasi-sentence unit is found. This will start at 1 with each new page (see below for instructions about natural sentences that span two pages).
- `paragrsno` - The serially numbered paragraph number for this page. This will start at 1 for each new page (see below for instructions about natural sentences that span two pages).
- `QScode` – The code assigned to the QS by the CMP coder, written in the margins of the document. Use "0" if it says "uncoded". It can happen that there are four digit codes. For manifestos from Eastern Europe, these should be recorded. In other countries, only the first 3 digits can be used (e.g. 107 for 1070, 408 for 4081). It may also happen (although it shouldn't according to the rules)[2] that a single QS receives more than one code, e.g. "503/402". In this case, record only the first code in `QScode`, but add both codes as written to `remarks`.

[1] Some of the manifestos actually lack these, which is at odds with the CMP coding rules.
[2] An exception are the "Europe" categories 108 and 110.

- `QSwordN` – The number of words in the QS. This should be as accurate as possible (but we recognize that some counting errors will occur when coding long texts). See below for details.
- `bullet` – An indicator variable with a value of 1 if this quasi-sentence was a bulleted point (i.e. starting with a bullet or a similar character) or if it is an item in a numbered list, or 0 otherwise. (If a bullet point consists of several QS, then this variable should be 1 for all of them.).
- `semicolon` – An indicator variable with a value of 1 if this quasi-sentence was part of a *natural sentence* ending with a semi-colon, or 0 otherwise. E.g. if we have one natural sentence with two quasi-sentences "[bullet point] We will do this, // and we will do that;", then the variable should be 1 for both quasi-sentences. In contrast "[bullet point] We will do this. And we will do that;" are two natural sentences and two quasi-sentences, and semicolon will be 1 only for the second natural sentence. The coding logic of this variable is therefore different from the one of `bullet.`
- `remarks` – Use this to include notes or remarks concerning the specific quasi-sentence.
- `parsingmarks` – Are quasi-sentences clearly delineated in the document? Assess this at the end of the coding process for the document as a whole (NOT necessary to do that quasi-sentence by quasi-sentence). Choose between a value of 1 when parsing marks are used *(almost) always*, 0 when used *(almost) never* and 0.5 when used *sometimes*.
- `codername` – Initials of who recorded the QS information.

## CODING INSTRUCTIONS

### Definition of a quasi-sentence:

The text unit identified as a quasi-sentence using bracket marks or lines in the coded text, by the CMP coder. (We do not second-guess CMP coders, rather we simply use the QS parsing given.)

### Definition of a *natural sentence*:

A *natural sentence* is a text unit delimited by: the following characters:
. ? ! ;
We note that ":" is not a delimiter.

For *bullet points*, we count each bulleted point as a NS, except the short text before the first bullet point. For example:

*Our party will do*

- *This.*
- *That.*
- *The other.*

consists of three natural sentences. (We leave the computerized implementation of bullet point parsing to be resolved later.)

If a single QS consists of several NSs we should consider the whole unit as a single NS. (According to the coding rules, this is not supposed to happen, but in some cases it clearly does.)

There are exceptional cases where the above rules fail to identify what can be considered a natural sentence. For instance, a bullet-point type list may not use bullet points or semi-colons but carriage returns to separate points (IS_1995_Awake_13523 page 1 bottom). Consider these as separate natural-sentences if it is obvious that they are separate statements.

**Definition of a paragraph:**

A paragraph ends with a hard return.

**When a unit spans more than one page**

If the last NS or QS or paragraph of a page overlaps the next page, then for the next page start counting at the first full NW/QS/paragraph to avoid counting twice. If an overlap occurs with regard to at least one of the units, then assign all component units to belong to the same page. Example: If q-s no. 40 within paragraph 23 on page 10 continues on page 11 and there are two further q-s within this paragraph no 23, then count them as belonging to paragraph 23 and also as q-s no 41 and no 42 on page 10, *rather than* as q-s no 1 and 2 on page 11). Reason: our experience shows it is much easier to being counting on each page (rather than for the document) when coding the quasi-sentences, because it is easier to correct mistakes. When processed for the final dataset, the coded serial numbers will be recoded to start at 1 for the document, not for the page. This recoding requires that the overlaps are handled as stated above.

**Counting words**

Count words and numbers, e.g. “Unemployment was higher than 2 million or 10% in 1989” would be considered as consisting of 10 words. “Unemployment was higher than 2 million or 10 per cent in 1989” would be considered as consisting of 12 words.

However, do not consider single characters separated by blank from numbers as separate: 10 %, 40 $, § 218 would be only one word each.

Words with hyphen count as one word (“mixed-member”), acronyms added in parentheses also count as one word.

Due to a lack of parsing marks, it may sometimes be unclear to which unit the introductory statements of lists belong. Reconsider the above example:

*Our party will do*

- *This.*
- *That*
- *The other.*

Without parsing marks in the text, the intro statement is counted as belonging to the first itemized point (only). So the word count here would be 5, 1, 2.

Since the cmp ignores headlines and sub-headlines, these do not belong to quasi-

sentences and their words should not be counted. Exception: Coder has clearly coded a headline as own QS.

**CODING PROCEDURE**

1. Identify the text by name in the catalog spreadsheet "`manifestos_wmarginalcodes_inventory.xls`".
2. Locate the scanned pdf file in the folder `Manifestos with margin codes_to do`, this will have a filename such as `AU_2001_NP_63810.pdf`, indicating that it is the Australian National Party manifesto from the 2001 election, coded as CMP party 63810.
3. Start a spreadsheet with the name `AU_2001_NP_63810.xls`, using the read-only template `co_year_par_xxxxx.xls`. Save it in the folder `recoding results`.
4. Code the quasi-sentences in the pdf manifesto using the instructions above. (It is probably easiest to work page by page. Start with typing in all codes from the margin of one page. Then add the information for those quasi-sentences. Then proceed with next page.)
5. Manifestos that are being coded and are not finished should be cut from the main text folder and pasted into the …in progress folder to rule out that somebody else starts coding them as well.
6. When done, save also a copy of the spreadsheet as tab-separated text file under the same name (only with suffix`.txt`) in the same folder.
7. Update the the catalog spreadsheet "`manifestos_wmarginalcodes_inventory.xls`" by marking the manifesto as "coded".
8. Move the coded pdf manifesto to the subfolder "`already recoded`" inside the "`manifesto texts`" folder.
9. Do the next one!

The data administrator (Thomas!) will take the results from each coded manifesto spreadsheet and append them to `consolidated_coding_results.dta`.