## Appendix A: Catalog of Constraints

This appendix collects the constraints used in this study in a format which should hopefully be useful in the future.

- It should be useful for guiding the articulation of constraints in future experiments on similar corpora.
- It should work as a starting point for a general catalog of constraints for diverse types of corpora.

The constraints are listed by broad category, and within each category, a number of simple operational approaches are listed.

### A.1 Constraint: Well-Defined Representation

Data should be represented in a well-defined form, preferably governed by a stringent and well-known standard (recommend UTF-8 for text, etc).

There should be no invalid code points, and no control characters except newline (U+000A) and possibly tab (U+0009) and carriage return (U+000D). If carriage returns are present, they should be consistently applied, one before each newline.

If other control characters are present, their intended semantics should be explicitly documented.

The significance of whitespace should be properly documented. There are two possibilities; either whitespace is sigificant, and each sequence of whitespace represents itself literally; or whitespace is insignificant, and can be trimmed without altering the semantics.

In modern text representations, nontrivial whitespace should perhaps be represented using some higher-level markup (see the next section).

### A.2 Constraint: Felicitous Markup

If metadata is present, it should be marked up using a standard, well defined notation (recommend XML for text).

For XML or SGML documents, a DTD should be provided, and the semantics of each element and attribute should be documented.

The same information should only be represented once. In other words, there should not be any overlap between the various elements and attributes of the DTD.

Where identifiers are used, they should obviously be unique and unambiguous.

### A.3 Constraint: Separate Metainformation

The markup should separate the samples from metainformation about each sample, and metainformation about groups of samples or the entire corpus.

Data in the actual samples which is useful as metainformation should be marked up so that it can be included or excluded as necessary.

For example, if samples contain the date of their publication as part of the sample itself, there should be a straightforward way to extract just the date for each sample (ideally, in machine-readable form), and usually also a way to extract each sample without the date information.

Similarly, samples should not contain boilerplate information. Each sample should represent itself, not some template. If boilerplate text is included, it should be similarly marked up for possible exclusion.

### A.4 Constraint: Duplicates are Identified

There may be legitimate duplicates in many corpora, but the corpus maintainer should indicate – by markup, sample naming, cataloging, or other means – which samples are duplicates, so that a practitioner can choose to exclude or subsample the groups of duplicates. Ideally, near-duplicates should also be similarly identified. They are frequently a more difficult source of problems than trivial duplicates.

### A.5 Constraint: Fidelity of Annotation

Markup should be consistently applied to all samples which meet the criterion for that markup. For example, if a language tag is present, a tag identifying a particular language should be present for all samples in that language, and no others. This generalizes to all types of annotations, such as syntactic tagging, topic tagging, etc.

### A.6 Constraint: Adherence to Selection Criteria

In order for a corpus to be useful, we have to know what it contains; if the contents are different from what the documentation states, they must be regarded as invalid.

This is obviously a complex task, but frequently, a well-defined content declaration can at least be used to superficially verify that the corpus does not contain invalid samples.

For example, in an email corpus, every sample should contain a message in a valid email format, be it RFC822 or some proprietary binary format. (Case in point: the SpamAssassin public corpus at http://spamassassin.apache.org/publiccorpus contains half a dozen files which are infrastructure, not actual samples.)

### A.7 Constraint: Documentation

This is a bit of a meta-constraint, and is not really conducive to automated verification at this point in time, but it is an important quality indicator.

A corpus should include documentation to help the practitioner identify the type and scope of the collection (ideally, without downloading it); and any conventions and standards used in the contents.

With the development of markup tools and formats, it is even conceivable that (at least parts of) the documentation itself could satisfy some rudimentary constraints that could be checked automatically.

## Appendix B: Constructing RCV2 Revision 1

This appendix contains brief instructions for constructing a revision of the RCV2 in accordance with the findings documented in this article.

1. Extract all the zip files.
2. Exclude all articles enumerated in on-line Appendix C. These are duplicates.
3. Correct language tags in articles listed in on-line Appendix D. Additionally, possibly remap some language codes (i.e. `jp` ⟹ `ja`, `zhtw` ⟹ `zh`).
4. Exclude all articles enumerated in on-line Appendix E. These are corrupted.

No renumbering is proposed; the resulting collection will simply have gaps in the article numbers where something was excluded.

Note also the 44 missing samples from the Latin American correction due to `zip` archive corruption.

Correction of country or topic tags in the XML markup has not been attempted. There are errors in these tags, but correcting them is a significant effort which is outside the scope of this article.

Similarly, detection of the actual language in each article has not been attempted. The corrections from Appendix D merely ensure that tags are consistent. Some articles contain multiple languages, and the present markup does not have an obvious, documented facility for marking up different languages within one article.

## Appendix C: Duplicate Articles (Online Appendix)

Due to the volume of the data (a list of 48,665 articles which should be excluded as identical duplicates of other articles in the collection), the text is only available as an on-line resource.

The collection of article identifiers is downloadable in machine-readable form from https://github.com/rcv2/rcv2r1/blob/master/online-appendix-c.txt

The process for identifying these duplicates is described in Section 3.4.

## Appendix D: Corrections for Language Tags

This appendix tabulates corrected values for the `xml:lang` element for 132 articles.

The information is also available on-line in machine-readable form from https://github.com/rcv2/rcv2r1/blob/master/online-appendix-d.txt

As discussed in the main article, the language tags cannot be trusted to reflect the actual language of the article's text. The changes outlined here merely ascertain that the articles are tagged consistently.

The language codes mostly adhere to ISO-639-1 except for `jp` and `zhtw`. (By current practice, the conventional BCP-47 markup would separate the language code and the region code by a dash; `zh-TW`. If language variants are supposed to be tagged, then perhaps the articles from the Latin American office ought to be tagged differently than the articles from the Spanish office, too; `es-AR`?)

The articles from the Japan office are mostly tagged as `jp` whereas the standard code for the Japanese language is `ja`. As detailed in the text, there are actually a few articles which have this language code; but for consistency, we revert that to `jp`. The opposite change might make more sense, but as discussed above, making all the language tags standards-compliant and consistent is a much larger effort in any event.

The remaining changes are for articles which have been tagged as English, even though the article is in fact in the expected default language for the subcollection.

| Article or Range | Corrected Tag |
|---|---|
| 83157 | `da` |
| 108676 | `sv` |
| 325931 | `de` |
| 341030 | `de` |
| 347498 | `de` |
| 466949-644953 | `jp` |
| 466966 | `jp` |
| 466970 | `jp` |
| 466974-467109 | `jp` |

The main article discusses this finding in Section 3.2.

## Appendix E: Fragmented and Corrupted Articles

This appendix contains a catalog of articles which are corrupted or fragmented, and which should thus be excluded as invalid samples.

The information is also available on-line in machine-readable form from https://github.com/rcv2/rcv2r1/blob/master/online-appendix-e.txt

This Appendix collects results from Sections 3.4.2, 3.5.2, and 4.2.1.

| Article | Article | Article | Article | Article | Article | Article | Article | Article |
|---|---|---|---|---|---|---|---|---|
| 2543 | 6346 | 7290 | 7654 | 7656 | 7895 | 7918 | 8531 | 8674 |
| 8795 | 8826 | 9118 | 9296 | 9351 | 9435 | 9521 | 9544 | 9547 |
| 9647 | 10130 | 10387 | 10441 | 10522 | 10589 | 10618 | 10721 | 10813 |
| 10819 | 10885 | 10886 | 11342 | 11458 | 11516 | 11675 | 12183 | 13450 |
| 13485 | 13982 | 15447 | 16408 | 16550 | 16625 | 49493 | 54940 | 55851 |
| 59634 | 61926 | 65773 | 66648 | 66770 | 66845 | 70447 | 76376 | 76833 |
| 76983 | 77156 | 77412 | 77615 | 77824 | 78400 | 79312 | 80480 | 80631 |
| 81100 | 81225 | 81501 | 81568 | 81977 | 82178 | 82239 | 82346 | 82394 |
| 82474 | 82574 | 82705 | 82806 | 82868 | 86292 | 88035 | 88432 | 89214 |
| 89215 | 90099 | 91322 | 92458 | 93964 | 96272 | 98320 | 98423 | 98554 |
| 98556 | 99070 | 99071 | 99324 | 99836 | 99940 | 100514 | 100520 | 101166 |
| 101324 | 101325 | 101551 | 101894 | 102032 | 102033 | 102288 | 103088 | 103507 |
| 103903 | 103936 | 103943 | 104101 | 104448 | 104823 | 105009 | 105201 | 105283 |
| 105490 | 105816 | 106078 | 106450 | 106852 | 107449 | 108342 | 108401 | 108403 |
| 108404 | 108890 | 109034 | 116888 | 122149 | 122903 | 194642 | 195692 | 196127 |
| 197560 | 197806 | 198034 | 198038 | 198145 | 198596 | 198792 | 198794 | 199020 |
| 199097 | 199181 | 199251 | 199619 | 201420 | 201508 | 202330 | 203842 | 204068 |
| 205108 | 205275 | 205735 | 207426 | 207670 | 207975 | 208182 | 208621 | 210977 |
| 214572 | 215391 | 218889 | 218014 | 220425 | 221106 | 222425 | 224034 | 224219 |
| 225667 | 226560 | 228751 | 230284 | 231153 | 234972 | 235579 | 236697 | 237909 |
| 237258 | 238156 | 240568 | 241647 | 241719 | 241904 | 241945 | 259072 | 275551 |
| 276514 | 276863 | 302224 | 308125 | 308704 | 309264 | 309793 | 309886 | 309947 |
| 310433 | 310434 | 310529 | 311363 | 311966 | 312301 | 312728 | 312762 | 314266 |
| 314335 | 314865 | 314866 | 316035 | 316406 | 316664 | 316952 | 316956 | 317154 |
| 317201 | 317571 | 317704 | 317889 | 318087 | 318325 | 319172 | 319416 | 319468 |

| Article | Article | Article | Article | Article | Article | Article | Article | Article |
|---|---|---|---|---|---|---|---|---|
| 319587 | 320586 | 321018 | 322679 | 323043 | 323752 | 324141 | 324752 | 324839 |
| 325286 | 325491 | 325926 | 328671 | 328821 | 329730 | 330799 | 331192 | 331257 |
| 332348 | 332582 | 333577 | 334290 | 334291 | 334966 | 335362 | 335447 | 335543 |
| 335716 | 336071 | 336399 | 336442 | 336714 | 336732 | 336999 | 337359 | 337693 |
| 338335 | 338770 | 339119 | 339525 | 340210 | 340775 | 341955 | 342042 | 343586 |
| 344376 | 345235 | 345620 | 350055 | 351495 | 351674 | 351697 | 351810 | 352709 |
| 352710 | 352712 | 353370 | 355311 | 355313 | 356549 | 356550 | 356819 | 356962 |
| 357104 | 357187 | 357307 | 358224 | 358838 | 359949 | 360334 | 360336 | 360897 |
| 360901 | 361825 | 362593 | 363822 | 363825 | 363828 | 364425 | 365471 | 365919 |
| 365922 | 366255 | 366436 | 366724 | 369255 | 369356 | 369388 | 369738 | 369739 |
| 369740 | 370077 | 370465 | 370780 | 371118 | 371404 | 371405 | 371666 | 372242 |
| 372243 | 373061 | 373280 | 373281 | 373925 | 373928 | 375172 | 375174 | 375175 |
| 375353 | 375600 | 375861 | 375862 | 376322 | 376865 | 377422 | 377488 | 399814 |
| 399932 | 399940 | 400259 | 400314 | 407387 | 415143 | 421802 | 421874 | 434061 |