ERRORS, OMISSIONS AND INCONSISTENCIES IN THE XML-VERSION OF THE BNC - Sebastian Hoffmann / Stefan Evert, July 2007

Note: The ordering of the sections reflects our evaluation of the seriousness of the problems – more serious issues are listed first.

1. DUPLICATE TEXTS AND PARTIAL DUPLICATES

The duplicate stretches of text listed below were found by comparing all s-units with 10 or 20 tokens in BNC-XML and printing out any repeated sequences to a file. The list below is the result of checking the more obvious cases. We suspect that there may be quite a few more duplicate stretches of text that could be detected with the help of a more sophisticated method (and more consistent checking). However, we would be surprised if any further texts were fully duplicated.

While we have not checked every single instance of these duplicates in earlier versions of the corpus, we suspect that almost all duplicates go back to BNC 1.0.

1.1. Spoken texts:

Identical or virtually identical texts and text passages:

- KNG and G55 (at least one <pause> tag is different)
- KNH and G56
- KNJ and G57 (one instance of <vocal desc="cough"/> in one text only)
- KNK and G58
- KB8:3509 onwards and KNS:103 to the end (s-unit 518): This is the same conversation, transcribed differently. There is no information about the speakers in KNS - but we have names etc. for the same speakers in KB8. It's quite interesting (and in fact rather discouraging) how significantly the two transcriptions differ!
- KDT:301-409 and KPW:905-1023 are identical. Once it's Robert and his mother and once it's Robert and an unknown person.
- KBN:1-345 is identical to KBN:346-690.

1.2 Written texts:

- File G3C.xml is a (near) duplicate of text HWX, while text G3C from the World Edition seems to be missing.
- EEE and HWK appear to be identical - mostly. The difference is that EEE has no s-unit 37 but jumps straight from 36 to 38 (without any other change in the text). Also, EEE has a reference to footnotes (<!--footnotes-->) after s-unit 87 which HWK does not have.
- HD1 and GXH are more or less the same text - but this seems to be a case where two people worked on the same text and marked it up slightly differently. The words are the same - but the formatting information isn't.
- CFL and C89 are identical with the exception of a table of contents and some minor formatting issues.

- HRH:444-1626 and HNV:1-1088 are again almost identical, with differences in formatting.
- CRP (complete text - 576 s-units) is the same text as FT7:1-734. (The difference in number of s-units is due to a section on events that is not part of the shorter text.)
- There are some problems with duplicate news items in two issues of the Guardian. This duplication occurs across different sections of the newspaper (in different BNC texts) - here are two instances we found, but we suspect that there may be more:
  - A95:33-124 (The Guardian, 1989-12-08, foreign news) is the same as A9E:48-139 (The Guardian, 1989-12-10, foreign news).
  - A99:1-49 (The Guardian, 1989-12-08, sports section) is the same as A9H:10-57 (The Guardian, 1989-12-10, sports section)
- This is not the only problem with Guardian texts:
  - A9S is a very short text (36 s-units) - it's from the Guardian edition of 1989-12-11: Foreign news pages. It consists of two news items. The first one (s-units 1-13) is the same as A9M:229-241 (same newspaper, same day, same section) and the second one (14-36) is the same as A9l:114-136 (same newspaper, same day, but **different section**: business)
  - AA1:1-84 (Guardian, 1989-12-13, **Foreign** news) is the same as A9W:743-827 (Guardian, 1989-12-13, **Home** news)
  - A7S:82-110 is the same text as A7S:112-140 (same text!), with a slightly different heading.
- While the Guardian texts appear to be particularly prone to this kind of error, the Daily Telegraph also shows some duplication:
  - AK6:1208-1254 is simply repeated again from 1255-1300.
- We also found a rather odd type of duplicated stretch of text in H0M:

  H0M:3147-3188 is replicated earlier in the same text (from 2402 to 2443). In fact, s-unit 3147 is appended to s-unit 2402 (The bold-faced part is s-unit 3147):

  2402 Jerome, the blue-jeaned bumboy with earring and dyed blonde **"Of course," he said, continuing to cut his food crisply,"you could have argued that the man was being exploited too.**

  The end of this "insertion" is in s-unit 2443, where parts of s-unit 3188 (in boldface) are cut off...

  2443 **She instanced her** of a long line of lagers, I consumed three Waistwatchers, two Seckburgers, an American Way and a double order of Tuckleberry Pie.

Evaluation of problem: **Serious**. We are surprised about the large number of duplicate stretches of text remaining in the BNC as we would have expected that this (known) issue would have been addressed in the third incarnation of the corpus.

2. MISSING <ALIGN>-TAGS:

There are 67,360 <align>-tags missing in the XML-version - they are the ones that are supposed to show overlap in spoken texts. A total of 236,179 align tags exist in BNC-XML - so 28.52% are now "orphans" in the sense that there's one <align>-tag but no corresponding counterpart. In the World Edition, this was not the case - only a total of 36 <ptr>-tags were orphaned. This basically means that in well over a quarter of all cases, users of the corpus will have no way to find out how long overlap lasts (or where it starts) for one of the speakers. There are also 8,802 instances where both instances of an <align>-tag have disappeared. There were 244,981 <ptr>-tag pairs in the World Edition as opposed to the 236,179 equivalent <align>-tags.

Evaluation: **serious**. Information that was carefully added to the corpus has been lost.

3. OTHER MISSING ITEMS:

The calculations given in the sections 3.1 to 3.4 rely on the tagUsage information given in the header. However, this information is apparently sometimes unreliable. As a result, these observations have to be treated with caution. For example:

In file JNW, we get <tagUsage gi="gap" occurs="9"/>. However, JNW has 16 <gap>-tags. The reason seems to be that some <events> in the World Edition have become <gaps> in the XML-version. There were 8 <gaps> in JNW in the World Edition.
Example of the changes:
s-unit 231:

> <event desc="gives code">
> becomes:
> <gap desc="code"/>

3.1 Missing pauses: 1,562 in total

Pauses appear to have been deleted by the conversion script when they immediately follow multi-word units.

File KRU:
XML-version:
<s n="710"><w c5="AV0" hw="home" pos="ADV">Home </w><**mw** c5="AV0"><w c5="PRP" hw="at" pos="PREP">at </w><w c5="ORD" hw="last" pos="ADJ">last </w></**mw**><w c5="PRP" hw="through" pos="PREP">through </w>

World Edition:
<s n="710"><w AV0>Home <w AV0>at last **<pause>** <w PRP>through

Evaluation: **serious**. Information that was carefully added to the corpus has been lost.

3.2. Missing unclear passages: 1062 instances

This again appears to be because of preceding multi-word units:

Text JSK:

World Edition:
<s n="215"><w PNP>He <w VVZ>goes <w AVP>on <w TO0>to <w VVI>say <vocal desc="clears throat"> <shift new=reading> **<w AV0>of course <unclear dur=13>** <shift> <event desc="microphone knocked"> <w AJ0>federal

XML-version:
<s n="215"><w c5="PNP" hw="he" pos="PRON">He </w><w c5="VVZ" hw="go" pos="VERB">goes </w><w c5="AVP" hw="on" pos="ADV">on </w><w c5="TO0" hw="to" pos="PREP">to </w><w c5="VVI" hw="say" pos="VERB">say </w><vocal desc="clears throat"/><shift new="reading"/><mw c5="AV0"><w c5="PRF" hw="of" pos="PREP">of </w><w c5="NN1" hw="course" pos="SUBST">course </w></mw><w c5="AJ0" hw="federal" pos="ADJ">federal </w>

Notice also how the second empty <shift>-tag has disappeared – so we now don't know that the speaker has stopped reading. See the following section.

Evaluation: **serious**. Information that was carefully added to the corpus has been lost.

3.3. Missing <shift>-tags: 176 instances

Again, multi-word units are apparently the culprit:

World Edition:
<s n="20"><w DT0>This <w AV0>then <w VBZ>is <w DTQ>what <w PNP>I <w VVB>command <w PNP>you<c PUN>, <w VVB-NN1>love **<w PNX>one another <shift>**<c PUN>.

XML-versin:
<s n="20"><w c5="DT0" hw="this" pos="ADJ">This </w><w c5="AV0" hw="then" pos="ADV">then </w><w c5="VBZ" hw="be" pos="VERB">is </w><w c5="DTQ" hw="what" pos="PRON">what </w><w c5="PNP" hw="i" pos="PRON">I </w><w c5="VVB" hw="command" pos="VERB">command </w><w c5="PNP" hw="you" pos="PRON">you</w><c c5="PUN">, </c><w c5="VVB-NN1" hw="love" pos="VERB">love </w>**<mw c5="PNX"><w c5="CRD" hw="one" pos="ADJ">one </w><w c5="DT0" hw="another" pos="ADJ">another</w></mw>**<c c5="PUN">.</c></s>

Evaluation: **serious**. Information that was carefully added to the corpus has been lost.

3.4. Missing gaps: probably at least 995 in spoken texts, 34,631 overall.

This one doesn't seem to be because of multi-word units – we have not been able to find any reason why they disappeared. This count is not very accurate because some <event>-tags have been converted to <gap>-tags. Also, some &formula; tokens have been converted into <gap>-tags. But still – something is not quite as it should be.

Text KS7:

World Edition:
<s n="1476"><w CJC>And <w AV0>now <w AT0>the <w AJ0>local <w NN2>results<c PUN>. **<gap desc="results">**

XML:
<w c5="CJC" hw="and" pos="CONJ">And </w><w c5="AV0" hw="now" pos="ADV">now </w><w c5="AT0" hw="the" pos="ART">the </w><w c5="AJ0" hw="local" pos="ADJ">local </w><w c5="NN2" hw="result" pos="SUBST">results</w><c c5="PUN">.

Another example:

World Edition:
<s n="5219"><w PNP>It<w VBZ>'s <w CRD>nine <w NN2>minutes <w PRP>to <w CRD>seven<c PUN>. **<gap desc=adverts>** <event desc="recorded jingle">

XML:
<s n="5219"><w c5="PNP" hw="it" pos="PRON">It</w><w c5="VBZ" hw="be" pos="VERB">'s </w><w c5="CRD" hw="nine" pos="ADJ">nine </w><w c5="NN2" hw="minute" pos="SUBST">minutes </w><w c5="PRP" hw="to" pos="PREP">to </w><w c5="CRD" hw="seven" pos="ADJ">seven</w><c c5="PUN">.</c><event desc="recorded jingle"/></s></u><u who="PS66C">

Evaluation: **serious**. Information that was carefully added to the corpus has been lost.

4 MISSING AND SUPERFLUOUS SPACES:

There are (tens of) thousands of missing spaces that were lost in the automatic conversion to XML. This is partly the case when additional highlighting is involved (e.g. italics):

File A0X – space missing before and after brackets:
<s n="1215">……<w c5="NP0" hw="peter" pos="SUBST">**Peter** </w><w c5="NP0" hw="guyett" pos="SUBST">**Guyett**</w><c c5="PUN">**,**</c><hi rend="it"><c c5="PUL">**(**</c><w c5="UNC" hw="ww/sept.89/p.803" pos="UNC">**WW/Sept.89/p.803**</w><c c5="PUR">**)**</c></hi><w c5="PRF" hw="of" pos="PREP">**of** </w>……

But here's one without additional mark-up intervening:

In World Edition:
<c PUQ>&bquo;<w NP0>Satan<c PUN>, <w AT0>the <w NN1>devil<c PUN>, <c PUL>(<w AT0>a <w AJ0>false <w NN1>accuser<c PUN>,

In XML-version:
<c c5="PUQ">‘</c><w c5="NP0" hw="satan" pos="SUBST">Satan</w><c c5="PUN">, </c><w c5="AT0" hw="the" pos="ART">the </w><w c5="NN1" hw="devil" pos="SUBST">devil</w><c c5="PUN">,</c><c c5="PUL">(</c><w c5="AT0" hw="a" pos="ART">a </w><w c5="AJ0" hw="false" pos="ADJ">false </w><w c5="NN1" hw="accuser" pos="SUBST">accuser</w><c c5="PUN">,

In contrast, there are also superfluous spaces, e.g. before a punctuation marker:

<w c5="NN1" hw="table" pos="SUBST">**table** </w><shift/><c c5="PUN">**!**</c>

The CQP conversion script tries to fix most of these errors, but in doing so introduces a small percentage of new errors. We believe that this is still better than the situation presented in the current version of BNC-XML.

Evaluation: For the general user, this is a **largely cosmetic** issue as long as the corpus tool correctly identifies the tokenization as indicated by XML-tags. However, depending on the tool and the task performed (e.g. on a version of the BNC which has been stripped of the XML-tags), we can imagine situations where pattern matches will fail to match because of the missing spaces. So this is **potentially serious**.

5. FURTHER FORMATTING ISSUES

- Some tokens (in particular <c> elements) contain a trailing newline (rather than just a blank), which makes handling them in line-based formats very inconvenient (but also for printing and other purposes).
- Whitespace between tokens is normally attached to the end of the first token. In some files (notably EFP through EX8), whitespace is attached to the beginning of (straight) quotation marks (") rather than the end of the preceding word.
- Some multiword tokens have not been split up correctly (into multiple tokens enclosed by a <mw> element), e.g. *in general* in file A07.
- Some <w>-units are just single blanks (presumably around 4400); in file EWW, there are 1472 such tokens. This seems to be a problem with mathematical formulas that immediately follow these empty w-units: <w c5="UNC" hw=" " pos="UNC"> </w><gap desc="formula"/>
- Most quotes have been normalised to Unicode characters LEFT SINGLE QUOTATION MARK and RIGHT SINGLE QUOTATION MARK, but there are plain ASCII double quotes (") in 368 texts. LEFT/RIGHT DOUBLE QUOTATION MARKs do not seem to be used at all. Similarly, “free” dashes are normally coded as EM DASH, but some 1% each appear as EN DASH or as plain ASCII - (these are typically mixed with EM DASHes in the same sentence to make the conversion bug more obvious).

Evaluation: **largely cosmetic**, but potentially more serious depending on type of inconsistency and context.

6. HEADER INFORMATION: SPOKEN TEXTS

Speaker IDs are supposed to be unique now (removing the old PS000 “unknown speaker”); but there are a few instances of PS001 left ("group of unknown speakers"), which also seem not to be declared in the file header. Here is a complete list of PS001 speakers:

Remaining PS001 speakers in the XML-version

filename	spid	speakerWords	speakerTurns
KB7	PS001	18	242
KD9	PS001	2	37
KDW	PS001	4	114
KDX	PS001	0	17
KR2	PS001	0	8

Sometimes speaker declarations have been forgotten – in fact, texts F7W and F8G simply have no speaker profile descriptions.

Here's a complete list of speakers without declaration:

F7WPSUNK [F7W], F7WPSUGP [F7W], F7XPSUNK [F7X], F7XPSUGP [F7X], F8GPSUGP [F8G], F8GPSUNK [F8G], HYGPSUNK [HYG], HYGPSUGP [HYG], J43PSUNK [J43], J43PSUGP [J43], J9FPSUNK [J9F], J9FPSUGP [J9F], J9GPSUNK [J9G], J9GPSUGP [J9G], J9HPSUNK [J9H], J9HPSUGP [J9H], J9JPSUNK [J9J], J9JPSUGP [J9J], J9KPSUNK [J9K], J9KPSUGP [J9K], J9LPSUNK [J9L], J9LPSUGP [J9L], JJDPSUNK [JJD], JJDPSUGP [JJD], JJEPSUNK [JJE], JJEPSUGP [JJE], JJFPSUNK [JJF], JJFPSUGP [JJF], JJTPSUNK [JJT], JJTPSUGP [JJT], JNLPSUNK [JNL], JNLPSUGP [JNL], JSKPSUNK [JSK], JSLPSUNK [JSL], JSLPSUGP [JSL], JSMPSUNK [JSM], JSMPSUGP [JSM], JT5PSUGP [JT5], JT5PSUNK [JT5], K66PSUNK [K66], K66PSUGP [K66], K78PSUNK [K78], K78PSUGP [K78], KB7PS001 [KB7], KD9PS001 [KD9], KDWPS001 [KDW], KDXPS001 [KDX], KJUPSUNK [KJU], KJVPSUNK [KJV], KLEPSUNK [KLE], KLFPSUNK [KLF], KR2PS001 [KR2]

We also remember having come across some speakers that are declared in the header but don't have any utterances in the body – we will compile a list of these in the next indexing process.

Evaluation: Given the low number of words involved, the PS001 issue is **of marginal significance**. Also, since all the speakers without declarations appear to be "unknown" anyway, their non-declaration in the header is a **cosmetic** issue, too.

7. S-UNIT NUMBERING:

There are about 15,000 instances where the s-unit numbering is not consecutive. In most cases, only one s-unit number is skipped, but there are also instances where the numbering jumps by a much larger number. It looks as this is the result of corrections made in BNC XML, e.g. by removing notes that had previously been erroneously part of the corpus. Example: in file KAY, s-unit 1 is missing but we instead have a comment on teacher's spelling - this was previously part of the text proper:

<wtext type="UNPUB"><!--
teacher's spelling corrections incorporated--> <head>

<s n="2"><w c5="VVB-NN1" hw="love" pos="VERB">LOVE </w><w c5="CJC" hw="and" pos="CONJ">AND </w><w c5="NN1" hw="marriage" pos="SUBST">MARRIAGE</w></s></head>

However, there also appear to be a number of inconsistencies as to the order of s-units. This was already the case in BNC-World; e.g. in text A06 where the numbering is frequently not consecutive in connection with stage directions:

<stage id=A06ST00F rend=it type=U>
<s n="**1065**"><w PNP>He <w VVZ>leans <w PRP>towards <w NP0>Wagner
</stage>
<s n="**1062**"><ptr target=A06ST00F><w VDB>Do <w PNP>you <w VVI>know <w DTQ>what <w PNP>I <w VVB>mean <w PRP>by <w AT0>a <w AV0>relatively <w AJ0>free <w NN1>press<c PUN>, <w NP0>Mr <w NP0>Wagner<c PUN>?<c PUN>…
<s n="**1063**"><w PNP>I <w VVB>mean <w AT0>a <w AJ0>free <w NN1>press <w DTQ>which <w VBZ>is <w VVN>edited <w PRP>by <w CRD>one <w PRF>of <w DPS>my <w NN2>relatives<c PUN>.

<stage rend=it type=U>
<s n="**1066**"><w PNP>He <w VVZ>throws <w AVP>back <w DPS>his <w NN1>head <w CJC>and <w VVZ>laughs<c PUN>.
</stage>
</sp>
<bibl>
<s n="**1064**"><w NN1>Act <w CRD>2
</bibl>

Evaluation: **Largely cosmetic**. However, if a corpus tool expects consecutive s-numbers (e.g. for displaying the larger context of a query hit), there might be problems. We recommend that this issue is raised in the documentation (if it isn't there already).

8. POS-TAGGING:

Some of the simplified wordclass tags are based on rather odd rules:

- DT0 is subsumed under "ADJ". The two most frequent adjectives in the BNC are therefore *this* and *that*. While it may be debatable whether words like *latter* or *former* are adjectives or not, we don't see any way that *this* or *that* could be classified as adjectives...
- The infinitive marker TO0 is subsumed under "PREP"
- ZZ0 is subsumed under "SUBST" - this is probably defendable, at least to some extent.

Evaluation: **not very serious, as long as users are made aware of this**. Depending on which corpus tool they use, users will be able to define their own simplified wordclass tags on the basis of the CLAWS C5 tag-set. Nevertheless, we are somewhat surprised that such a counter-intuitive grouping of C5-tags was chosen.

9. DOCUMENTATION ERRATA

- The documentation wrongly suggests that ambiguity tags made up of two different wordclasses all receive an "UNC"-value in the simplified wordclass tag. In reality, they receive a tag according to the first of the two tags.
- In the documentation (Section 9.8), the c5-tag "POS" is subsumed under punctuation. In the corpus, however, all "POS"-tags have an "UNC"-tag in the simplified wordclass.