
Mitton, Roger; Hardcastle, David and Pedler, Jenny (2007) *BNC! Handle with care! Spelling and tagging errors in the BNC.* Presented at the Fourth Corpus Linguistics Conference, 27-30 July 2007, Birmingham, U.K.

BNC! Handle with care!
Spelling and tagging errors in the BNC

Roger Mitton, David Hardcastle and Jenny Pedler
School of Computer Science and Information Systems,
Birkbeck, University of London
R.Mitton@dcs.bbk.ac.uk

"You loose your no-claims bonus," instead of, "You lose your no-claims bonus," is an example of a real-word spelling error. One way to enable a spellchecker to detect such errors is to prime it with information about likely features of the context for *loose* (as a verb) as compared with *lose*. To this end, we extracted all the examples of *loose* used as a verb from the BNC (British National Corpus, XML edition, written part (Burnard, 2007)).

There were, apparently, 159 occurrences of *loose* (VVB or VVI). However, on inspection, well over half of these were not verbs at all (tagging errors) and over half of the rest were misspellings of *lose*; far from providing us with useful information for correcting the *loose*-for-*lose* error, they were themselves examples of it. Only about fifteen percent of the 159 occurrences were genuine occurrences of *loose* as a verb. This prompted us to undertake a small investigation into errors in the BNC.

Let us make it clear at the outset that we consider the BNC a very valuable resource, that we make great use of it and that we are most appreciative of the work that has gone into it. We are not "knocking" the BNC or the efforts of its creators. But since many people may be tempted to use it as a sort of gold standard of correct English for training software or for education, it may be salutary to share with others some of our findings regarding its imperfections.

1. Spelling errors

We began by drawing up a list of about 4,000 pairs of words that resembled each other, in spelling or pronunciation or both, such that one might be written in mistake for the other, such as *accept* and *except* or *accursed* and *accused*. With the aid of the Xaira query software (Xaira version 1.23, 2007) and using only the written part of the BNC, we went through this list, having a quick look at the concordance for one of the members of each pair to see how often it appeared in place of the other. To make the task tractable, we confined our attention to those pairs where errors accounted for at least ten percent of the occurrences – a quick scan through the first page or two of a concordance sufficed to show whether further analysis was worthwhile.

The texts in the BNC are mostly publications, and many will have been edited and proofread, so we did not expect to find a large number of errors. Nonetheless we ended up with a list of about seventy. Table 1 presents some of the examples where the errors outnumber the genuine occurrences. (All occur in a variety of texts; occurrences as proper nouns are excluded.)

Word	**N of genuine occurrences**	**Target word(s) (with frequency)**
withe	0	with (15)
calender	0	calendar (14)
ail	2	all (54)
tor	3	for (65)
canvasses	4	canvases (14)
posses	5	possess (10)

polices	7	policies (13)
abut	8	about (30)
wold	10	would (17), wild (3), world (3)
loosing	10	losing (21)*
rime	25	time (29)
* Includes seven from a single source – the Leeds United email list		

Table 1: A selection of real-word errors from the BNC

The table shows that, for example, there were fifty-six occurrences of the word *ail* in the written part of the BNC, excluding its use as a proper noun or acronym, but only two of these were genuine occurrences; the other fifty-four were misspellings of the word *all*.

Some of these errors will have originated with the creators of the text – the Leeds United email list, for example, was a particularly fruitful source. Others may have crept in during the process of transferring printed material into electronic form – *ail* for *all*, *tor* for *for* and *rime* for *time* all look like OCR errors.

A common thread runs through those in the table and most of the others in the full list. If a relatively rare word resembles a much more common one, an occurrence of the rare one is likely to be an error. *Withe* (a pliable twig) and *calender* (a machine for smoothing cloth or paper) provide extreme examples of this; both appear in the BNC only as errors (though the two occurrences of *calenders* (plural) are both correct). *Fiat* (common noun) for *flat*, *minster* for *minister* and *manger* for *manager* are other (less extreme) examples from the full list.

That our list contains so many of these is partly an artefact of our method. Fourteen occurrences of *calender* jump off the screen when every single one is a misspelling. By contrast, if *calender* had been a much more common word, fourteen errors, or even forty, would have been buried in hundreds of correct uses and would not have been noticed.

Nonetheless, the pattern itself is genuine and is of some interest for spellchecking. An earlier study of the misspellings of less frequent words (Damerau and Mays, 1989) concludes that, when a less frequent word occurs, it is much more likely to be a correct spelling than a misspelling of some other word. As a consequence they recommend the use of large dictionaries for spellchecking, though they suggest that special treatment may be needed for very rare words. Our findings add a refinement to this: it's not just the rarity of the rare word that needs to be taken into account, it's also the commonness of the word that it resembles.

A further complication arises when orthography is related to part-of-speech, as in *practise* (verb) versus *practice* (noun), or *affect* (verb) versus *effect* (noun), especially in the latter case since *affect* can, rarely, be a noun and *effect* can, occasionally, be a verb. Not surprisingly, confusion over the spelling of these words caused some problems for the tagger. Table 2 presents some results for *practise* and *affect* (errors on *practice* and *effect* were much less frequent).

practise(s) correctly used as a verb	1184
practise(s) used as a noun instead of *practice(s)*	133
affect(s) correctly used as a verb	5723
affect(s) correctly used as a noun	19
affect(s) used as a noun instead of *effect(s)*	95

Table 2: Errors on practise(s) and affect(s)

The words *ti* and *depute* are two further examples of how figures from the BNC cannot always be taken at face value. The word *ti* is, surprisingly, in the dictionary, as the seventh note of the tonic solfa scale (do-re-mi), and, leaving aside proper nouns and acronyms, it occurs 167 times in the written part of the BNC. Not one of these, however, has anything to do with do-re-mi. About half are renderings of the word *to* spoken in a Yorkshire accent (all from one source), and the rest are divided between mistypings of *it, to* and *time (ti me),* and assorted oddities.

The word *depute* provides an example of how the BNC can be misleading even when it is correct. *Depute* is an ordinary English verb, though an uncommon one in that form (more often *deputed*, one would think), yet it occurs eighty-five times to *deputed*'s twenty-three. On inspecting the concordance it appears at first sight that these are all errors – the required word is clearly *deputy* – but in fact they are not misspellings but variant spellings. *Depute* is an old spelling of *deputy*, which is still current in Scotland. – all the examples are from Scottish sources.

2. Tagging errors

Our approach to tagging errors was different. In the course of enriching a dictionary with frequency data from the BNC (Mitton, 1986, Pedler, 2003) and, later, developing a lemmatizer (Hardcastle, 2007), we kept running across tagging errors. Of course, since the program that was used to do the tagging (CLAWS4 (Garside and Smith, 1997)) had a probabilistic component, a certain proportion of tagging errors are to be expected, and great efforts have been made, successfully, to reduce the incidence of these (Smith, 1997, Fligelstone et al. 1997). But the errors that we encountered were on perfectly straightforward words that the tagger persistently mistagged. We noticed them simply because the tags that CLAWS repeatedly gave them were different from the tags that they had in our dictionary.

Not all of these discrepancies were due to errors in the BNC. Sometimes the dictionary's tags were incomplete, often because, being based on a publication from the 1970s, they were out of date; the word *bin*, for example, was listed in the dictionary only as a noun, whereas it is sometimes tagged in the BNC, correctly, as a verb. But in many cases, the tags that CLAWS had given to these words were simply wrong. The following lists present a selection of these. The numbers in brackets are the frequencies in the written part of the BNC, excluding proper nouns; words with a frequency less than ten are excluded. Where a word was often but not always mistagged in the way described, some of its other tags might have been correct, but not necessarily; *retrograde*, for example, was mostly tagged as a noun, occasionally as a verb, but never as an adjective.

Adjectives often or always mistagged as nouns:

retrograde (175), moribund (114), open-mouthed (109), outbound (109), faraway (87), politic (80), pectoral (66), bespoke (65), taciturn (60), lumbar (58), deadpan (57), disconsolate (50), dicey (50), unbidden (47), workaday (46), lank (46), downtrodden (45), inbound (43), dinky (40), aquiline (40), hale (39), bedridden (39), bonkers (38), akimbo (35), fleecy (33), elfin (31), unisex (30), inclement (30), underfloor (27), shipboard (27), isosceles (27), spick (24), conjoint (20), superfine (16), houseproud (16), hangdog (15), svelte (14), foursquare (14), slapdash (13), prolix (13), alfresco (13), standoffish (12), oversea (12), hirsute (12), gimcrack (12), footsore (12), drear (12), bounden (12), nonstick (11), gaga (11), way-out (10), gluey (10)

Adverbs often or always mistagged as nouns:

ergo (36), ahoy (26), out-of-doors (25), pronto (23), side-saddle (20), agin (17), abeam (16), e'er (15), edgeways (13), leastways (12), molto (11), overarm (10), midships (10)

Verbs often or always mistagged as nouns:

mown (73), cajole (50), revamp (39), enshrine (37), sublet (36), crash-landed (36), unblock (35), hanker (33), oversaw (30), rehouse (28), fester (28), undervalue (27), peels (27), overshoot (27), exhale (26), loiter (25), countersunk (25), foist (23), saith (21), foment (21), outmanoeuvre (20), misspelt (19), etch (17), abridge (16), suss (15), sicken (15), bombards (14), quoth (13), jack-knifed (13), redone (12), overbalance (12), inverts (12), drool (12), sickens (11), misrepresents (11), jut (11), burgeon (11), blaspheme (11), behead (10)

Words often or always mistagged as adjectives:

turn-off (73), relent (48), lift-off (40), misrepresent (36), turmeric (31), sled (25), monosyllable (18), flowerbed (18), nitty-gritty (16), brush-off (16), volute (13), tizzy (12), sheepfold (10), disproof (10), chivvy (10), biped (10)

Words often or always mistagged as comparative or superlative adjectives:

haulier (150), plunger (30), natter (25), glazier (24), lounger (22), camper (15), outlier (13), gondolier (12), pannier (12)

second-best (46), ingest (32), headrest (14)

Words often or always mistagged as adverbs:

fortnightly (113), unseemly (90), neighbourly (43), half-yearly (35), dally (30), measly (28), niggardly (22), matronly (21), fleshly (15), drizzly (15), squally (13), pally (13), googly (12), maidenly (11), twiddly (10), doily (10)

Words often or always mistagged as verbs:

centigrade (90), effendi (78), unbeliever (36), unbelievers (29), centipede (25), lounger (22), shibboleth (16), athwart (15), aether (15), derring-do (14), wether (13), get-togethers (13), adipose (13), salsify (12), loungers (11), wheresoever (10), groovers (10)

Singulars often or always mistagged as plurals, or vice-versa:

politeness (232), gens (92), confetti (62), mews (54), kudos (54), scabies (41), portcullis (28), corgi (24), rickets (23), ravioli (23), patchouli (22), balls-up (21), bathos (19), albumen (18), mumps (16), reredos (14), brae (13), kohlrabi (12), pyrites (11)

woodlice (40), narcissi (27), kibbutzim (15), corgis (14), syndics (11), levis (10)

Why did CLAWS make these mistakes? Why did it even consider, for example, that the noun *haulier* might be a comparative adjective? It did not have any problem with *hauliers*. Why did it consider "adverb" to be the only possible tag for *fortnightly*? (It sometimes does function as an adverb, of course, but more often it's an adjective.) The answer presumably lies in its procedure for assigning candidate tags in the first place. CLAWS was designed as a robust tagger – it would produce tags for any set of words given to it. Since no dictionary could be expected to contain all the strings that the tagger might encounter, its dictionary was supplemented with a set of heuristics for guessing at the tag(s) of an unknown word. Presumably the words in the above list were simply missing from its dictionary, so it guessed at a tag or set of tags for them, and got it wrong or only partly right. *Haulier* looks like a comparative adjective, along the lines of *livelier* and *sillier*, whereas *hauliers* looks more obviously like a plural noun. *Fortnightly* looks like, and can be, an adverb, but less obviously like an adjective.

3. Should corpus errors be corrected?

How should corpus errors be dealt with? Assuming for a moment that funds were available to pay someone to do it, would we want the errors to be simply corrected?

Presumably, for errors that have crept in during the processing of the corpus, from the conversion of the original source texts into electronic form and on through the subsequent tagging (POS and otherwise), the answer is yes – we would prefer to have them corrected. There is no obvious virtue in preserving OCR errors or CLAWS mistaggings.

For errors in the source texts themselves, the answer is not so clear. It seems likely that many users of the BNC, such as teachers of English using it as source material or developers of language-processing software using it as training data, would prefer to have an error-free corpus. On the other hand, the corpus is a record of what English text looked like at a particular time, and perhaps the errors are part of that. It is not impossible that someone might want to compare the incidence of errors in the BNC with that in other corpora. A researcher in the future, for example, wishing to compare the level of errors in the newspapers of the day with the levels in the late twentieth century, would want a faithful record of the originals, not a cleaned up version.

It may also be argued that, for scientific purposes, there is virtue in stability. If someone develops a program and reports certain results from running it over the BNC, someone else should be able to replicate their work. But this would only be possible if the BNC had not changed in the meantime.

But do we have to choose one or the other? Can we not have our cake and eat it? Given that the BNC is tagged, we don't need to correct the errors, just mark them, preferably with the suggested correction. The errors would be preserved for anyone who was interested in them, but users who would prefer correct text could read the corrections instead. This practice was already adopted to some extent during the creation of the corpus (Burnard, 2000 : 12). Could it not be continued?

The kind of correcting effort that we have in mind would not be a one-off major project, but rather an ongoing minor one. Many people use the BNC and must notice errors all the time. If they sent them in to a BNC maintenance unit somewhere (Oxford?), perhaps in a standard form via the web, someone could glance through them when sufficient had accumulated, adjudicate on any dubious ones and add the necessary error tags to the corpus. A new version of the BNC would be released from time to time. No doubt this would be a lot more complicated in practice than we are making it sound – Burnard (2000:12), who is in a position to know, likens the folly of such an enterprise to the Walrus's scheme for sweeping sand off the beach – but it looks to us as though it ought to be possible and, in our opinion, worthwhile.

Acknowledgements

Our thanks to George Mitton for inspecting about 4,000 concordances and listing those that showed a high incidence of real-word spelling errors.

References

Burnard, L. (2000) 'Where did we go wrong? A retrospective look at the British National Corpus', in Kettemann B. and G. Marko (eds), Language and Computers, Teaching and Learning by doing Corpus Analysis. Proceedings of the Fourth International Conference on Teaching and Language Corpora, pp. 51-70, Rodopi

Burnard, L. (2007) Reference guide for the British National Corpus XML edition http://www.natcorp.ox.ac.uk/XMLedition/URG/

Damerau, F.J. and E. Mays (1989) 'An examination of undetected typing errors', *Information Processing and Management 25 (6)* : 659-64

Fligelstone, S., M. Pacey and P. Rayson (1997) 'How to generalize the task of annotation', in R. Garside, G. Leech and T McEnery (eds) Corpus Annotation, pp. 122-36, Addison Wesley Longman

Garside, R. and N. Smith (1997) 'A hybrid grammatical tagger: CLAWS4', in R. Garside, G. Leech and T McEnery (eds) Corpus Annotation, pp. 102-21, Addison Wesley Longman

Hardcastle, D. (2007) Lemma and inflection tables available at http://www.davidhardcastle.net

Mitton, R. (1986) 'A partial dictionary of English in computer-usable form', *Literary and Linguistic Computing 1 (4)* : 214-5

Pedler, J. (2003) 'A corpus-based update of a 'computer-usable' dictionary', *Proceedings of the Eighth International Symposium on Social Communication* : 487-92

Smith, N. (1997) 'Improving a tagger', in R. Garside, G. Leech and T McEnery (eds) Corpus Annotation, pp. 137-50, Addison Wesley Longman

Xaira online documentation (2007) http://www.oucs.ox.ac.uk/rts/xaira/