Spoken BNC2014 EAS vs. the demographic part of the BNC: What can a study of tag questions tell us?

Karin Axelsson (University of Gothenburg, Sweden)

The study of English spoken conversation has benefited tremendously from the demographic part of the *British National Corpus* (BNC) launched in 1994 (Burnard 2007). However, language change is, as always, going on so a corpus of spoken English reflecting the language of the early 2010s might give us a better understanding of how spoken English has changed over twenty years. This is the background to the development of *Spoken BNC2014* (Love *et al*. 2017 fc.) to be launched in 2017. However, some researchers have already been allowed access to a subset. This *Early Access Subset* (EAS) has about 5 million tokens and is comparable in size to the demographic part of the original BNC. Henceforth, the EAS will be referred to as *BNC2014E* and the demographic part as *BNC1994D.*

BNC2014E appears to be suitable for the study of one of the most typical features of English conversation: *canonical tags*, as in *It's going to rain, isn't it?*, including non-standard *innit* derived from *isn't it/ain't it* (Andersen 2001:106). The term *declarative tag question* (DecTQ) is here used for the combination of a declarative *anchor* (e.g. *It's going to rain*) and a *tag* (*isn't it*, *innit,* etc.).

The aim of the present study of DecTQs in BNC2014E is to investigate their frequencies, formal features, functions and sociolinguistic features in contemporary English and make comparisons to their use in BNC1994D. After considerable random thinning of the search result in BNC2014E, 497 DecTQs were identified and analysed. Separate searches, without thinning of the results, were also made for the non-standard tags *innit* and *ain't it* as well as for TQs with imperative anchors (ImpTQs). For BNC1994D, data from my doctoral dissertation (Axelsson 2011) was mainly used (1,315 instances) but some additional analysis of data from BNC1994D was performed for the present study, particularly as to sociolinguistic features. One hypothesis was that the tag *innit* is on the increase and that standard canonical DecTQs might therefore be on the decrease. There are indeed fewer DecTQs in BNC2014E than in BNC1994D (2,795 vs. 5,062 pmw) (without *innit*: 2,737 vs. 4,623 pmw), but the frequency of *innit* (including *in it*) has dropped even more: from 439 pmw in BNC1994D to just 72 pmw in BNC2014E (in the separate search for all instances of *innit*). These surprising results made me look closer at the comparability of the two corpora for the study of TQs.

Firstly, the two corpora have been compiled in different ways. BNC2014E uses crowd-sourcing where the requirements of good recording quality seem to favour more focused conversations than in BNC1994D, where the randomly selected respondents were told to record all spoken interactions during two days, i.e. also when the conversation itself was not the main activity going on. This difference is reflected in the fact that there are virtually no second-person ImpTQs (as in *don't tell her will you*) in BNC2014E (1.4 pmw), i.e. much less frequent than in BNC1994D (40 pmw). ImpTQs deal with exchanging goods and services and not exchange of information (Axelsson 2011) and may thus tend to be more common if other activities than just talking are going on. The more focused conversations in BNC2014E may also result in fewer misunderstandings. The proportion of constant-polarity DecTQs (as in *she's working is she*) is just five per cent in BNC2014E (vs.

almost ten per cent in BNC1994D). One reason to use constant-polarity DecTQs is to seek verification (Kimps 2007), e.g. if one questions what one believes to have heard another interlocutor say. Another consequence of the more focused conversations is probably a slightly higher level of style. An indication of this is that the proportion of reversed-polarity DecTQs with ellipsis of the subject and/or the finite in the anchor (as in *typical English weather isn't it*) is clearly less common in BNC2014E (9%) than in BNC1994D (17.5%).

Secondly, BNC2014E is less well balanced for sociolinguistic features than BNC1994D, where the respondents were selected to reflect all ages, social groups and regions proportionally. As to the age of the speakers in BNC2014E, the category 19–29 is clearly overrepresented, whereas speakers younger than that are underrepresented. The fact that children below ten are almost absent in BNC2014E may also affect the language used by adults, as they would probably tend to talk differently to children than to adults. This may be an additional reason why second-person ImpTQs are rare in BNC2014E compared to BNC1994D: instances such as *Don't touch anything will you* appear only in BNC1994D. The social grade groups with the highest normalised frequencies of DecTQs in BNC1994D are C2 and DE (significantly higher than for the rest of the corpus, i.e. grades AB, C1 and "unknown"). In BNC2014E, grades C2 and D are poorly represented (only about 2% each), whereas grades A, B and E contribute 24–31 per cent each. The facts that BNC2014E is not well balanced for sociolinguistic features, that the social grades of 38 per cent of the speakers in BNC1994D are unknown, and that social grade might affect the use of TQs complicate the comparison between the two corpora.

Thirdly, there are differences in the transcription principles. The low number of *innit* in BNC2014E might partly be due to the transcription guidelines saying "only use *innit* when you are sure: otherwise either use *isn't it* or *ain't it*" in combination with the clearer recordings required. Another problem is that the transcribers for BNC2014E were not allowed to use very much punctuation, practically only question marks, and that the size of the corpus is given in tokens (including quotation marks) instead of just words. For the calculations of frequencies and formal features above, the size of the two BNC corpora have been recalculated to exclude punctuation (BNC2014E then has 8.7 per cent more words than BNC1994D: 4,707,081 vs. 4,329,797). However, tokens are used in the calculations for sociolinguistics features as the sociolinguistic metadata is only supplied in tokens.

Fourthly, although there are as many as 376 different speakers represented in BNC2014E, some individual speakers have fairly large shares of the whole corpus. All the normalised frequencies presented above are actually calculated based on data from a subcorpus of BNC2014E, where all the utterances of 14 very prolific speakers (those with more than 75,000 tokens in the corpus) are excluded; in this subcorpus, there are 238 DecTQs uttered by 119 different speakers. The reason for this reduced dataset is that some speakers seemed to skew the results severely, in particular one speaker (an old man from Norfolk), who uses TQs to such an extent that, if the whole BNC2014E had been considered, his share would have been extremely high: 9.8 per cent of all DecTQs, 22.2 per cent of all instances of the tag *innit* and 45.1 per cent of all instances of the tag *ain't it*. Among the 14 excluded speakers, the normalised frequencies of DecTQs range between 554 and 18,658 per million tokens. This very large variation in individual frequencies makes it vital for the study of TQs

that corpus material is spread over many individuals and that no individual is allowed to have a high share of the words/tokens.

It has been suggested to me that, instead of comparing frequencies of TQs per tokens/words, it would be more adequate to use frequencies per utterances, sentences/s-units, clauses or finite verbs. The number of utterances is similar in the two corpora (just one per cent more in BNC1994D). The number of s-units is given for BNC1994D but not for BNC2014E. The number of clauses is difficult to calculate in both corpora; however, it is possible to compare the frequencies of finite verbs, which is related to the number of finite clauses. The frequency of finite verbs (imperatives excluded) is about five per cent lower in BNC2014E, so calculations per finite verb would decrease the difference in the frequency of TQs somewhat between the two corpora. Interestingly, there are notable differences as to the distribution of finite verbs: finite forms of *be* and *do* are more common in BNC2014E than in BNC1994D, whereas it is the other way around for finite forms of *have*, modals and lexical verbs. In the case of *have*, this is reflected in a significantly lower proportion of DecTQs with forms of *have* in the tag in BNC2014E than in BNC1994D (the only statistically significant difference between the proportions of tag subjects and tag verbs between the two datasets).

The requirements of good recordings and the detailed transcription guidelines make BNC2014 relatively easy to use. The contexts of examples are much more coherent than in BNC1994D. The analysis whether a match is actually a TQ is facilitated by the fact that change of speaker is indicated already in the concordance lines with speaker codes, which also offer quick links to the sociolinguistic information. Despite the comparability problems described above, the endeavour to compile a corpus of spoken conversation from the early 2010s with comprehensive sociolinguistic information must be praised. Hopefully, the final version of Spoken BNC2014 will be somewhat better balanced for sociolinguistic features.

References

Andersen, G. (2001). Pragmatic markers and sociolinguistic variation: a relevance-theoretic approach to the language of adolescents. Amsterdam: Benjamins.

Axelsson, K. (2011). *Tag questions in fiction dialogue*. (Doctoral dissertation). University of Gothenburg, Göteborg. Retrieved from <http://hdl.handle.net/2077/24047>.

Burnard, L. (2007). *Reference Guide for the British National Corpus (XML Edition).* Retrieved from http://www.natcorp.ox.ac.uk/XMLedition/URG/.

Kimps, D. (2007). Declarative constant polarity tag questions: a data-driven analysis of their form, meaning and attitudinal uses. *Journal of Pragmatics*, 39(2), 207–291.

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017 fc). The Spoken BNC2014: designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, 22(3).