TGAW22
Britishness Test for Written Corpus Texts

Dominic Dunlop

Version 1.0 10th February 1993

1 Background

The British National Corpus is intended to be a corpus of modern British English. The selection procedures for written texts to be considered for inclusion in the corpus (see [**?**], [**?**]) are intended to discriminate in favour of British authors, but do not exclude non-British authors. This paper describes an empirical test intended to exclude from the corpus any written work for which it can be established that a majority of the authors have no clear connection with the UK.

This paper does not discuss spoken texts, the selection criteria for which are set out in [**?**].

2 Applicability

The test described in the next section should be applied to any written text, or part of a written text, considered for inclusion in the corpus, provided that the text, or part of a text, appears to be 5,000 words or more in length[1].

The intention is to exclude ephemeral items, newspaper and magazine articles from the testing process, while still applying it, for example, to individual papers in conference proceedings or journals, and to individual short stories in anthologies. It is judged too time-consuming to apply the test to each ephemeral item, newspaper or magazine article in the corpus, and, in any event, we usually have little or no information about the authors of such pieces.

[1] There is no requirement that an accurate word-count is obtained before deciding whether or not to apply the test: an estimate is adequate.

3 The Test

After looking at material on a book's dust jacket, in its preface, introduction and acknowledgements (and the corresponding parts of works having sections attributed to particular authors), and at any other sources that are to hand (library catalogues, general knowledge etc.), sequentially apply the following critera, relative to the time at which the work was completed[2], **stopping as soon as a decision one way or the other has been reached**:

1. If the sole author, or half or more of the joint authors, can be described as UK residents, the work is acceptable, no matter what the nationality of the author or authors.

2. If the sole author, or half or more of the joint authors, can be described as UK citizens or as UK expatriates, the work is acceptable, no matter where the author or authors are domiciled.

3. If the sole author, or half or more of the joint authors, appear to have been in the UK for a period of two years or more[3] immediately before the work was completed, the work is acceptable, no matter what the nationality or domicile of the author or authors.

4. If the sole author, or half or more of the joint authors can be described as having been born in the UK, the work is acceptable, no matter the nationality or domicile of the author or authors, **unless ??** applies.

5. If the sole author, or half or more of the joint authors, appear to have been in some other country for a period of two years or more immediately before the work was completed, the work is **not** acceptable.

6. If the spelling in the text does not follow British conventions[4], the work is **not** acceptable.

7. If the country of first publication is not the UK, the work is **not** acceptable.

8. If a work cannot be excluded from the corpus as a result of questions 1–**??** having been answered, or having been skipped because of insufficient information, the work is taken to be acceptable. That is, in the absence of evidence to the contrary, an author is assumed to be British.

[2] In the case of published works, *completed* may be taken to mean "submitted for publication in something close to its final form". The process of publication may take a year or more, during which the author or authors generally provide little input. There is therefore no requirement that authors be in the UK as a work moves from submission to publication.

[3] This period is intended to respresent the minimum time in which a course of post-graduate study may be undertaken, or the period of employment after which a position may be said to be permanent.

[4] Note in particular that *Center*, *Defense*, *Theater* ... may appear in the names of places or organizations in texts which otherwise follow British spelling conventions.

4 examples

4.1 *Biting at the Grave*, Padraig O'Malley

1. *Is author a UK resident?*
   No: author teaches in Boston
2. *Is author UK citizen or ex-pat?*
   Don't know: no information
3. *Was author in UK for 2 or more years while writing the book?*
   No: although he visited UK, many US sources are credited as well.
4. *Was author born in UK?*
   Don't know. Could be from Northern Ireland; more probably from the south.
5. *Was author outside UK for 2 or more years before writing the book?*
   Yes.

Following answers not relevant to acceptability

6. *British spelling?*
   (No: US)
7. *First published outside UK?*
   (Yes: USA.)
8. *Answers to all of 1–**??** too unclear to allow decision?*
   (No: definite answer to point **??**.)

This is a definite failure at point **??**. It could be made into a pass if an affirmative answer to **??** or **??** could be established.

4.2 *Doubt*, O S Guinness

1. *Is author a UK resident?*
   No: author lives in Washington DC
2. *Is author UK citizen or ex-pat?*
   Don't know: no information.
3. *Was author in UK for 2 or more years while writing the book?*
   Probably not. (The author did study in England, but some time before writing the book.)
4. *Was author born in UK?*
   No: China.

5. *Was author outside UK for 2 or more years before writing the book?*
   Probably.

6. *British spelling?*
   Yes.

7. *First published outside UK?*
   No: simultaneous Australian, US and UK publication.

8. *Answers to all of 1–**??** too unclear to allow decision?*
   Yes.

According to the rules, this scrapes in, because there is insufficient information by the time the last question is reached.

4.3 *TA Today, A New Description of Transactional Analysis*, Ian Stewart & Vann Joines

1. *Is author a UK resident?*
   One of the two authors may be, one may not be.

2. *Is author UK citizen or ex-pat?*
   Don't know: no information

3. *Was author in UK for 2 or more years while writing the book?*
   Don't know: no information

4. *Was author born in UK?*
   Don't know. No information.

5. *Was author outside UK for 2 or more years before writing the book?*
   Don't know. No information.

6. *British spelling?*
   No: US.

Following answers not relevant to acceptability

7. *First published outside UK?*
   (No: Simultaneous UK and USA.)

8. *Answers to all of 1–**??** too unclear to allow decision?*
   (No: definite answer to point **??**.)

This text is not acceptable on the grounds of spelling: we know too little about the authors to be able to decide whether either or both of them have UK connections. If a positve answer for any of questions 1–**??** could be established for either author, the text would become acceptable despite its spelling conventions.

References

[BNCW08] Written Corpus Design Specification

[TGAW14] Spoken Corpus Design Specification

[TGAP21] Guidelines on Random Text Selection