Genre coherence and distinctiveness in the International Corpus of English: A quantitative approach

Axel Bohmann, Albert-Ludwigs-Universität Freiburg

This paper introduces an innovative method for exploring relationships among sub-groups in a corpus of linguistic data. The specific focus is on the coherence and distinctiveness of text categories in ten national sub-corpora of the International Corpus of English (ICE) project (Greenbaum & Nelson 1996), with the aim of proposing a typology of genres. ICE corpora comprise four spoken and eight written text categories, for each of which the following relationships are explored:

- How linguistically distinct is this text category from the other corpus texts on the whole?
- Within the text category, how distinct are texts representing different varieties of English?
- For text categories with more fine-grained sub-distinctions, how distinct are the sub-categories from each other?

In order to quantify linguistic overlap/distinction, the study relies on the ten dimensions of variation developed in Bohmann (2019). These express general textual properties and have been constructed empirically on the basis of co-variances among 276 individual linguistic variables. Relationships between groups of texts in this ten-dimensional space are expressed via the Bhattacharyya coefficient (Bhattacharyya 1943), a measure of the overlap between two multivariate distributions. For instance, in response to the first research question above, it is possible to calculate the Bhattacharyya coefficient for overlap between all spontaneous conversation (S1A) text samples and all remaining text samples.

Addressing the three research questions above by means of the Bhattacharyya coefficient results in a categorization of corpus text types into three broad classes:

- Highly coherent text categories that are clearly distinct from other categories and show moderate regional distinction: Creative writing, Private dialogue.
- Less coherent categories that show relatively strong cross-varietal distinctiveness: Non-professional writing, Correspondence, Reportage, Instructional writing, Persuasive writing.
- Less coherent categories with lower degrees of cross-varietal distinctiveness: Public dialogue, Unscripted monologue, Scripted Monologue, Academic writing, Popular informational writing.

The findings are relevant both at a theoretical and a methodological level. In terms of the former, the importance of genre in mediating linguistic variation has long been recognized (e.g. Hundt & Mair 1999), but the relationship among genres themselves has received less attention. The typology offered above may help corpus compilers establish and empirically verify appropriate levels of granularity in their sampling frame; it may also help researchers identify the kinds of text in which regional divergences may be expected, the points of overlap between genres that enable the gradual spread of a feature from one to the other, etc.

At the methodological level, use of the Bhattacharyya coefficient is not limited to specific corpus linguistic questions. It can express overlap between any two groups along any number of dimensions, such as vowel formant measurements or frequencies of specific lexical items. As such, it is useful for studies of group relationships in sociolinguistics and dialectology more generally. Its ability to incorporate multiple dimensions of variation at once is promising for holistic perspectives on linguistic distinctiveness in the spirit of dialectometry (e.g Szmrecsanyi 2013).

References

Bhattacharyya, A. 1943. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*. 35: 99–109.

Bohmann, Axel. 2019. *Variation in English Worldwide: Registers and Global Varieties.* Cambridge: Cambridge University Press.

Greenbaum, Sidney & Gerald Nelson. 1996. The International Corpus of English (ICE) project. *World Englishes* 15(1). 3–15.

Hundt, Marianne & Christian Mair. 1999. "Agile" and "uptight" genres: The corpus-based approach to language change in progress. *International Journal of Corpus Linguistics* 4(2). 221–242.

Szmrecsanyi, Benedikt. 2013. *Grammatical Variation in British English Dialects: A Study in Corpus-Based Dialectometry.* Cambridge: Cambridge University Press.