

A corpus-based approach to investigating twenty-first-century prescriptivism

Morana Lukac (Leiden University, The Netherlands)

1 Introduction

The publication and the remarkable success of Lynne Truss's prescriptive guide to punctuation *Eats, Shoots and Leaves* (2003) led Beal to observe that prescriptivism had returned with a vengeance in the twenty-first century (2010). Whereas the number of publications of similar books containing prescriptive usage advice is continuously on the rise (Tieken-Boon van Ostade, in progress), not all linguists share Beal's pessimistic view on the current quality of public debates on language use. Crystal, for instance, claims that at the moment we are "moving away from an institutionalized prescriptivism towards a more egalitarian linguistic era" (2006), and Cameron suggests that the web may be a platform facilitating the interaction between lay people and linguists (2012) and consequently popularizing the descriptive approach to language. Empirical research on the topic, however, is lacking, rendering it difficult to establish easily which between the opposing arguments is more accurate.

In my analysis, I explore a corpus of metalinguistic comments comprising three internet registers – blogs, Wikipedia pages and below-the-line comments of online articles – and one pre-internet register – letter-to-the-editor (LTE) sections of newspapers by way of multidimensional analysis (Biber, 1988). Wikipedia talk pages hosts discussions on grammar and usage that result in creating and editing Wikipedia entries. The participants in the remaining three genres simply comment on and express their views on usage and grammar. The analysis reveals that significant variation exists regarding how usage is discussed in the analysed registers, and it can largely be explained by the existence of different communities of practice (cf. Wenger, 1999; Eckert, 2006) in which the metalinguistic comments are embedded. Wikipedia pages on which usage and grammar are discussed, the results suggest, provides a space for forming an especially close-knit community of practice. In the second part of the analysis, I compare – by means of key semantic domain analysis – Wikipedia pages that describe problematical aspects of grammar (as an example of a type of present-day online sources) with a sub-corpus of entries compiled from a diachronic database of usage advice literature published between 1770 and 2010, called the Hyper Usage Guide of English or the HUGE database (Straaijer, 2014). The results of this analysis reveal, on the one hand, that Wikipedia entries manage to obtain a high level of objectivity and avoid prescriptive accounts due to the editors' commitment to the community principles as opposed to traditional usage guides, which contain many *ipse dixit* judgments. On the other hand, the results also show that we can identify continuity throughout the history of prescriptivism – as Beal and Tieken-Boon van Ostade suggest – with regard to the topics that

are discussed in usage debates.

2 The corpus and analysis

The research initiative Bridging the Unbridgeable: linguists, prescriptivists and the general public at Leiden University has traced the history of prescriptivism in the English language with the help of the HUGE database, which comprises 77 usage guides published from 1770 onwards. Next to usage guides, the liveliest arenas for discussions on usage today can be found online. In the present study, I explore these arenas within a self-compiled corpus of metalinguistic comments comprising three internet registers – blogs, wikis and below-the-line comments of online articles – and one pre-internet register – letter-to-the-editor (LTE) sections of newspapers – in all of which members of the public express their views concerning usage.

In building the corpus of metalinguistic comments, I implemented Biber’s proposal for representativeness in corpus design (1993), in which a set of objective criteria is introduced for determining the number of texts that need to be included in corpus analysis. . The underlying idea of Biber’s proposal was: the more variation, the more texts are needed to represent a genre, and the less variation, the fewer texts. Although Biber’s aim was to achieve objectivity, some amount of subjectivity cannot be avoided in setting up a study. I needed to thus arbitrarily decide on a number of texts that I would include in the final corpus: I established that the number would amount to a total of 1000 texts, which is approximately twice the size of the corpus Biber compiled in his 1988 study (57 (481 texts), since the texts in the genres I included tended to be much shorter than the ones from the original study. This procedure led me to establish my final corpus design shown in Table 1.

Table 1. The corpus of metalinguistic comments

Genre	Blog	Wiki	BTL	LTE	Total
Texts / Words	6,441 / 107	15,850 / 181	27,343 / 238	131,455 / 474	181,089 / 1,000

In the next step, I allocated the four registers to Biber’s six dimensions with the help of the Multidimensional Analysis Tagger (Nini, 2015). After that, I set out to identify the extent to which the four registers were similar to each other and to the 23 registers introduced by Biber (1988) by performing a statistical analysis called hierarchical clustering; in the analysis I considered the six dimensions jointly (see Berber Sardinha, 2014). The results of the hierarchical clustering analysis are reported on in Figure 1.

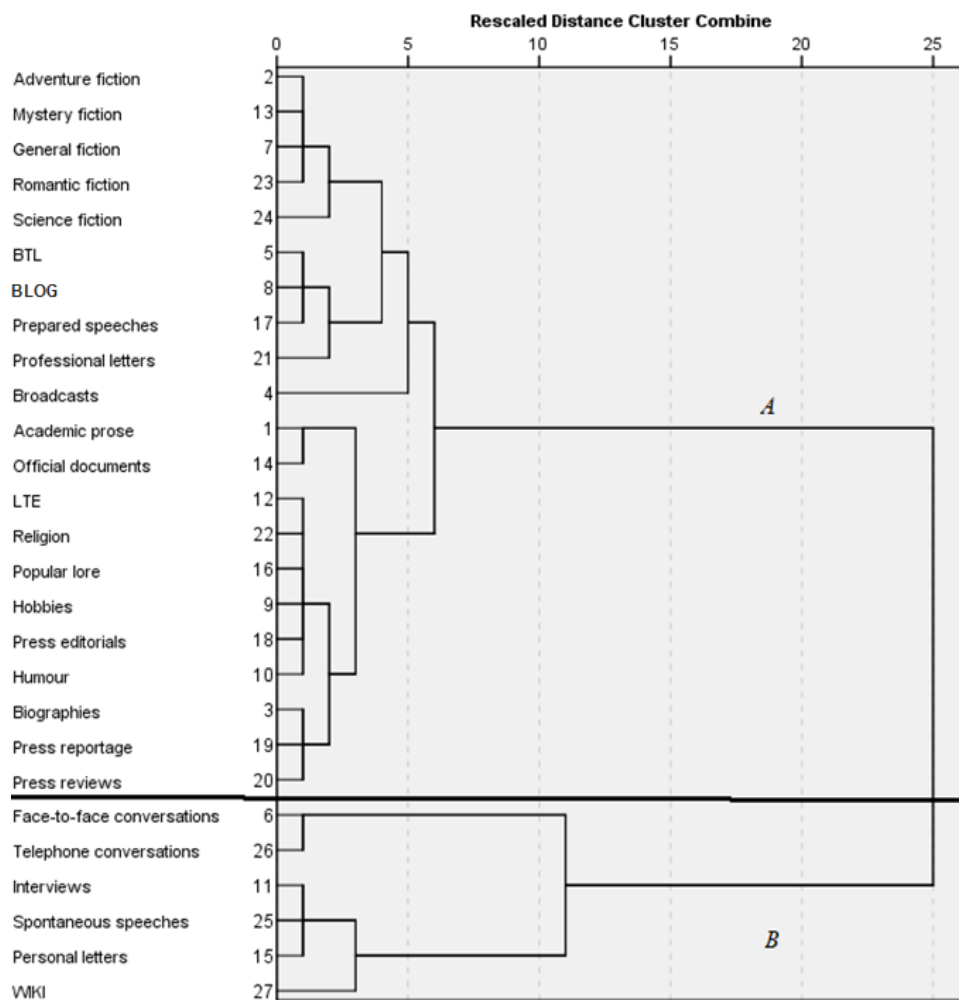


Figure 1. Hierarchical clustering of genres

Much of what we can observe from the general bipartite division in the tree diagram in Figure 1 indicates a distinction between spoken and written genres. The blog and BTL comments are clustered together being the most similar among the genres I analysed, and they are also closely linked to the interactive formal or written modes of communication. Letters to the editor, the only pre-internet genre among the four, are grouped together with discursive and informational genres, and they are also linked to other newspaper genres. The most distinct here are wikis, which belong to cluster B, and are grouped, on the lowest level in the cluster, together with spontaneous speeches and personal letters.

If anything, the results in Figure 1 indicate that the four registers, and consequently, the usage discussions found in them are heterogeneous. A closer examination reveals that the differences among them depend mostly on the existence of a close-knit community of practice and the communicative goals of their members. Wikis – websites that are created to enable collaborative projects

– have stronger ties and are more practice-oriented than the other registers.

3 Comparing twenty-first-century to traditional usage advice

With the introduction of Web 2.0 the discourse on language use entered the new media enabling re-contextualization of the previously established one-way communication between the advice-seeking lay community and the advice-giving language experts. I compare these two contexts by looking at the Wikipedia entries and their secondary Wikipedia Talk pages on the topics related to usage and the entries in usage guides retrieved from the HUGE database. For this I used the web-based Wmatrix tool (Rayson, 2009) through which I identified key words, and subsequently key semantic domains in the two corpora that would help determine significant differences between the Wikipedia-type entries and those found in usage guides.

The results show that Wikipedia entries obtain a higher level of objectivity and avoid prescriptive accounts as opposed to usage guides. Wikipedia entries are a product of collaboration as opposed to usage guides, which are usually works of individual authors. Due to this fact, as well as the guidance of Wikipedia principles, balanced discussions on Talk pages and regular edits, Wikipedia entries on language use largely reflect critical, up-to-date accounts relying primarily on linguistics as a discipline and actual usage, instead of single authorities and traditional gate keepers. The actual impact of Wikipedia and other forms of social media on usage and their popularity when compared to other sources of advice on usage is yet to be explored.

References

- Beal, J. (2010). The grocer's apostrophe: popular prescriptivism in the 21st century. *English Today*, 26(2), pp. 57–64.
- Berber Sardinha, T. (2014). 25 years later: Comparing Internet and pre-Internet registers. In T. Berber Sardinha & M. Veirano Pinto (Eds.), *Multidimensional Analysis, 25 years A tribute to Douglas Biber* (pp. 81– 108). Amsterdam and Philadelphia: John Benjamins Company.
- Biber, D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing*, 8(4), pp. 243–57.
- Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.
- Cameron, D. (2012). *Verbal Hygiene: The Politics of Language* (Rev. Ed.). London and New York: Routledge.
- Crystal, D. (2006). Into the Twenty-First Century. In L. Mugglestone (Ed.) *The Oxford History of English* (pp. 394–413). Oxford: Oxford University Press.
- Eckert, P. (2006). Communities of practice. In K. Brown (Ed.) *Encyclopedia of Language and Linguistics* (2nd Ed.) (pp. 683–5). Amsterdam: Elsevier.
- Nini, A. (2015). *Multidimensional Analysis Tagger* (Version 1.3).

- <https://sites.google.com/site/multidimensionaltagger/>.
- Rayson, P. (2009). *Wmatrix: a web-based corpus processing environment*. Computing Department, Lancaster University.
<http://ucrel.lancs.ac.uk/wmatrix/>.
- Straaijer, R. (2014). *Hyper Usage Guide of English database*. <http://huge.ullet.net/>.
- Tieken-Boon van Ostade, I. (in progress). *English Usage Guides: The Biography of a Genre*.
- Truss, L. (2003). *Eats, Shoots & Leaves: the Zero Tolerance Approach to Punctuation*. London: Profile Books.
- Wenger, E. (1999). *Communities of Practice: Learning, Meaning, and Identity*. Cambridge: Cambridge University Press.