FILOLÓGIA.HU • XVI. évfolyam (2025) 1–2. szám 88–92. • 10.59648/filologia.2025.1-2.6

Dávid Veljanovszki

Separating the Grain from the Chaff: Relying on Linguistic Clues to Tell Real News Apart from Fake News

Book Review

Grieve, Jack – Woodfield, Helena 2023: *The Language of Fake News.* Cambridge: Cambridge University Press. Online: https://doi.org/10.1017/9781009349161

Despite the best of intentions and a sense of reasonable caution, one is constantly exposed to the danger of manipulation by untrue information supplied wholesale by a seemingly endless and ever expanding range of media outlets these days. One of the most natural ways of safeguarding against an unwanted onslaught of such information could be limiting one's media consumption to trusted content providers. However, as it happens in a number of instances, even the most widely renowned platforms and titles may occasionally not spare their readers the nuisance of unwittingly becoming victims of misinformation or – even worse – of disinformation. Although the average reader of newspapers, printed or electronic, cannot be expected to have the time or the capacity to analyse the linguistic features of news texts to establish their veracity, a handful of empirically based findings about the apparently systematic differences between real news and fake news might furnish the alert news consumer with some useful points of reference.

The Language of Fake News by Jack Grieve and Helena Woodfield offers a detailed report on a corpus-based project analysing the journalistic products of Jayson Blair, a former *New York Times* reporter, widely known for his track record of publishing fake news during the early 2000s. Choosing Blair's texts from his notorious period and compiling a corpus out of them are explained by the authors in terms of three factors: topic variation, the status of the texts within the sample as fake news and time. While acknowledging the limitations of the small sample of 64 articles (*viz.* a few shorter articles were excluded due to their inadequacy for yielding measures for the relative frequency of grammatical forms), the researchers were thus in a good position to control for variables that could otherwise have detracted from validity: register, dialect, authorship, news outlet, topic and political bias – the absence of the last of these resulting from the relative temporal remoteness of the news stories reported on.

In an effort to help the reader better contextualise the subject under investigation, as well as to identify a clear research focus, Grieve and Woodfield devote much of the introductory section of their book to the controversies surrounding the

definition of fake news, embedded in a detailed overview of the history of deception over a period of four centuries. Based on historical examples and taking account of the various media and genres for transmitting fake news ranging from 16th- and 17th-century *avvisi,* through war-time fake news and Soviet-era *dezinformatsiya* to the latest conspiracy theories chiefly propagated on platforms outside the mainstream (e.g. blogging and social media), they make a fine distinction between fake news and false news, referring to the author's express intent to deceive as the underlying motivation for the former. The authors point to the disadvantages of employing natural language processing models, citing their excessive emphasis on classification accuracy at the expense of explaining patterns of language use and giving precedence to variation in language content over variation in language structure. Thus, designating the need for a linguistic description of fake news, concentrating on stylistic variation as a niche area, as opposed to simply plotting topic trends, the researchers settle on an empirical framework informed by functional theories of language use drawing on register variation (cf. Biber–Conrad 2019), as well as by a principled distinction between misinformation and disinformation (cf. Rubin 2019). Grieve and Woodfield also stress the special relevance of their undertaking in terms of contributing to forensic deception detection and investigative linguistics by meeting the general legal requirement of using theory-driven tools to generate evidence admissible in court.

In Section 2 of the book, this research orientation is further supported by an extensive review and well-placed criticism of previous research on the language of fake news. While recognising the merits of the existing, largely veracity-based corpora of fake news (e.g. LIAR or the fake news database of BuzzFeed), the authors expose numerous weaknesses of these collections, blaming their reduction of the true–false distinction to a single binary variable, their lack of attention to so-called partial texts (i.e. individual texts containing both real and fake news), the inevitable politicisation of labelling owing to the involvement of external judges (e.g. fact-checking organisations and mainstream media organisations) in the vetting process and the failure to control for decisive variables such as register, authorship and dialect. In line with their empirical goals, the researchers advocate a new framework grounded in theories of disinformation (i.e. intended to deceive in matters of societal consequence, cf. Stahl 2006 and Rubin 2019) and register variation. Combining these two theoretical strands, Grieve and Woodfield predicate their experiential basis on Biber's model of multidimensional analysis (Biber–Conrad 2019) and argue that disinformation should be treated as a form of register.

After discussing details of the journalistic career and track record of fraud of Jayson Blair in Section 3, the authors proceed to present the corpus in Section 4. In describing the criteria for the selection of the research subject, as well as for the inclusion of texts in the corpus, Grieve and Woodfield highlight favourable conditions for studying examples of fake news where the journalist sought to deceive his readers into believing information he himself did not believe was true (cf. Type II Fake News in Grieve–Woodfield 2023: 13), along with the opportunity to control for register, dialect, authorship, news outlet, topic and political bias. Accordingly, the analysis was conducted along the lines of three dimensions: topic, status as fake news and time. The quantitative results of the investigation utilising the Multidimensional Analysis

Tagger (Nini 2019) are presented and interpreted in Section 5. The analytical framework worked with normalised values for 67 grammatical features based on multidimensional register analysis (cf. Biber 1988). For 28 features, the results are visualised in the form of boxplots, with general communicative purpose in the English language, and in Blair's writing, juxtaposed for each feature. The choice of the format and layout is particularly helpful from the point of view of appreciating two aspects considered simultaneously for each grammatical feature: register analysis and the fake news – real news dichotomy. It is also praiseworthy that, with a view to controlling for topic variation as a variable, an additional subset of textual data was created based on a single fake news story, the D.C. Sniper Attacks, where no major differences were found in the relative frequency of the grammatical features under analysis compared to the rest of the sample featuring a variety of topics.

Apart from presenting and discussing the quantitative results on the relative frequency of the selected grammatical forms in the sample, Section 5 also exploits the dataset for two other variables: information density and conviction. In their treatment of the former, the authors appear to equate information density with concision and succinctness of expression manifesting itself mostly in real news specimens through features such as nominalisation, the use of the gerund, present participle post-nominals, past participle post-nominals, the infinitive with *to* and the use of time adverbials. Conversely, the fake news specimens within the sample tended to be marked by a relatively higher proportion of arrangements such as structures involving the use of the pronoun *it,* 3rd person pronouns, demonstrative pronouns, 1st and 2nd person pronouns, predicative adjectives, attributive adjectives, emphatics, downtoners, place adverbials, split auxiliaries, subordinators and conditional subordination – syntactic features mostly associated with spoken registers. While the prevalent patterns in terms of register are clearly established, it would be interesting to learn whether the overrepresentation of linguistic features more characteristic of spoken registers in fake news texts was simply the result of a slapdash attitude on the journalist's part in his desperate effort to cobble together news reports out of a patchwork of dubious or outright falsified sources, or whether lower levels of information density may be linked to specific types of fraud, like plagiarism, pretence, false claims or systematic fraud, as identified by the *New York Times* investigation. This would, however, require closer scrutiny of the circumstances of the news stories in question.

In conjunction with the final target of the corpus-based study, conviction, it was demonstrated that real news texts contained a relatively larger number of linguistic exponents typically associated with taking a stance and conveying a sense of evidentiality. These included suasive verbs (e.g. *allow, demand, determine, insist,* etc.), possibility modals, by-passives (i.e. passive constructions with an agent) and public verbs (i.e. speech act verbs). At the same time, fake news items displayed more diverse verbal forms, presumably due to the higher concentration of relative clauses introduced by *wh-* relative words.

Given the novelty of the empirical angle adopted, the thorough corpus-based investigation conducted, as well as the authors' cognisance of the limitations imposed by the size and scope of the sample, Grieve and Woodfield's work represents an important and revelatory contribution to our understanding of the nature of fake

news, to forensic discourse analysis as a field of academic inquiry, as well as to the study of media discourse at large. Despite their initial scepticism about the possibility of discerning systematic differences between real news and fake news, the authors have succeeded in identifying systematic linguistic variation between the two datasets, with considerable implications for investigative linguistics. Nonetheless, much as the researchers' rationale for choosing a journalist whose activities and the events he reported on are devoid of any sense of immediate relevance and hence of the risk of politicisation thanks to the distance in time between the present and the early 2000s may be appreciated, one cannot help asking if similar systematic patterns could also be observed for genres that have arisen since, such as blogs, tweets, social media posts or clickbait-motivated short online articles changing in content by the minute. Therefore, future research projects could seek to find out if real news and fake news also systematically differ across these genres.

References

Biber, Douglas 1988: *Variation across Speech and Writing.* Cambridge: Cambridge University Press. Online: https://doi.org/10.1017/CBO9780511621024

Biber, Douglas – Conrad, Susan 2019: *Register, Genre, and Style.* Cambridge: Cambridge University Press. Online: https://doi.org/10.1017/9781108686136

Grieve, Jack – Woodfield, Helena (2023): *The Language of Fake News.* Cambridge: Cambridge University Press. Online: https://doi.org/10.1017/9781009349161

Nini, Andrea 2019: The Multi-dimensional Analysis Tagger. In: Sardinha, T. Berber – Pinto, M. Veirano (eds.): *Multi-Dimensional Analysis. Research Methods and Current Issues.* London: Bloomsbury Academic. 67–94. Online: https://doi.org/10.5040/9781350023857.0012

Rubin, Victoria L. (2019): Disinformation and Misinformation Triangle: A Conceptual Model for 'Fake News' Epidemic, Causal Factors and Interventions. *Journal of Documentation* 75/5: 1013–1034. Online: https://doi.org/10.1108/JD-12-2018-0209

Stahl, Bernd Carsten 2006: On the Difference or Equality of Information, Misinformation, and Disinformation: A Critical Research Perspective. *Informing Science: The International Journal of an Emerging Transdiscipline* 9: 83–96. Online: https://doi.org/10.28945/473

Dávid Veljanovszki, Ph.D., is a linguist, translator and conference interpreter in multiple language pairs. He has been a University Lecturer of English Applied Linguistics since 2001. His areas of expertise include ELT methodology, research methodology in language pedagogy, varieties of English, English as a lingua franca and the history of the English language. He has widely published on various aspects of discourse analysis in academic settings. Currently, he is Associate Professor at the Department of English and German Studies of Ludovika University of Public Service, Budapest. E-mail: veljanovszki.david@uni-nke.hu

Dávid Veljanovszki 2025: Separating the Grain from the Chaff: Relying on Linguistic Clues to Tell Real News Apart from Fake News. Book Review. *Filológia.hu* 2025/1–2: 88–92.