

CORPUS RELATED DIGITAL HUMANITIES: Interdisciplinary Micro Perspectives

Edited by:

Mădălina CHITEZ, Roxana ROGOBETE, Loredana BERCUCI,
Andreea DINCĂ, Raluca-Andreea PETRUȘ



**Editura Universității de Vest din Timișoara,
2022**

General Editors' Preface / 7

Introduction / 9

Talking Their Way to Success: How Do YouTube Beauty Vloggers Speak? / 15

Andreea Dincă

How Material Is Poetry? A Micro-Stylistic and Quantitative Analysis / 21

Alexandru Foitoş

The Language of Romanian Administration: An Interview-Based Corpus Case Study / 25

Elena-Mirabela Grama

Harlem: Photographs of Change / 31

Adela-Diana Chindriş

Linguistic Features of Billboard Hits / 42

Roxana-Delia Prohăp

In Their Own Words: A Study of The Language of Hollywood / 49

Eliza Bătrîn

Persuaded into Beauty / 56

Florina-Maria Andercău

Depiction of Diseases in Poetry: Plague, Cholera, Covid-19. A Corpus-Based Study / 63

Margareta-Luisa Boicu

How Did Elisabeth Cady Stanton's 'Tone' Change

Over a Period Of 10 Years (1887-1897)? / 72

Teodora Bogoevici

The Language of Teenagers: Where Does "Gen" Come From? / 79

Bianca-Marinela Socaci

Linguistics of Horror: Corpus-Based Analysis of Horror Films / 89

Karla-Claudia Csűrös

Digital Humanities and Research for British and American Studies / 98

Loredana Bercuci

The Language of Gamers: What Do Critics Investigate and What Does

That Tell Us About Video Games? / 102

Ciprian-Victor Meteş

Notes on Editors / 114

Notes on Authors / 115

At the Intersection of Corpus Linguistics and Digital Humanities

The establishment of Corpus Linguistics (CL) as a discipline (Tognini-Bonelli, 2001) or cross-disciplinary research method (Brezina & McEnery, 2020), depending on the theoretical perspective, has been associated with the emergence of computer processing techniques that facilitated the transformation of language related resources (e.g. texts, books, speech transcripts) into searchable databases (Dash, 2009). A similar development has taken place in the case of Digital Humanities (DH), a domain that came into being along with the digitization of the first documents relevant for the humanities studies. Both fields have undergone conceptual and terminological transformations due to changes in research and implementation focus, but, ultimately, they have reached a level of recognition within the research and professional community. The transition from Roberto Busa's "mechanolinguistics" (Johansson, 2008, p. 34) towards corpus linguistics, as well as the step from "humanities computing" towards Digital Humanities have been marked via dissemination landmarks: launching of the well-established journal *Computers and the Humanities* in the 1960s, respectively Blackwell Publishing's volume *A Companion to Digital Humanities* in the 2000s (Brügger, 2016).

There are, however, significant differences between the two domains that emerged from the same necessity: to adapt traditional approaches in humanities to the newest technological developments. While corpus linguists are interested in collecting large amounts of linguistic data for statistics-based analyses relevant for various linguistics sub-disciplines such as lexicology, applied linguistics, discourse analysis, computational linguistics or any domain that investigates "linguistic variation" (Reppen et al., 2002), digital humanities experts' primary roles is the digital preservation, display and popularization of cultural heritage assets. At the same time, while Corpus Linguistics has opened up important pathways for research and applications in various language related humanities disciplines (e.g. literature and translation studies, language pedagogy), Digital Humanities has a much broader approach spectrum since "DH unites traditional humanities disciplines such as archaeology, history, philosophy, linguistics, literature, art, music and cultural studies with computing tools and techniques, e.g., hypertext/media,

data visualisation, IR, statistics, 3D modelling, data/text mining, digital mapping and spatial analysis” (Golub et al., 2021, p. 4). On the hand, the two disciplinary areas, that have long been embedded in separate research communities, have started to manifest affinities for one another. This is reflected in the amount of studies that incorporate both domains within a shared framework (e.g. Santos, 2019; Hinrichs et al, 2019; Zottola, 2020). What appears to be the common ground is the focus on language related analyses and the intersection of such studies with other disciplines, such as health, historical, and political studies (Curza, 2012), psychology (Roivainen, 2020), sociology or education (O’Keeffe & McCarthy, 2010).

Interdisciplinary micro perspectives

The topic of the first volume in the *SPOR* series is *Corpus Related Digital Humanities: Interdisciplinary Micro Perspectives*. The book is about the use of corpus linguistics and corpus related methods to conduct research on topics specific to the field of humanities, thus emphasizing transdisciplinary synergies between the conventional and the technology supported approaches. Each chapter reflects the authors’ perspective on a particular component of a disciplinary field, such as general linguistics, literature or cultural studies. As the research expertise profile of the authors varies from novice to advanced researchers, we believe that having different academic perspectives on similar methods might shed light on their applicability and research replicability potential.

This represents, in a nutshell, the primary agenda of the book. What we want to achieve with this book and what should take this book apart from others, having a similar thematic focus (e.g. McGillivray & Tóth, 2020), is the didactisation of the selected topics: each research or theoretical case study is presented in such a way that the reader receives all the information necessary (e.g. type of data, specific terminology) to understand the components of the research process.

The single prerequisite for obtaining a complete picture of the “micro perspective” illustrated by each chapter is familiarization with corpus specific data processing. The junior researchers (i.e. CODHUS student assistants and interns), who have authored several chapters in this volume, have been trained to process linguistic data in corpus format and to use various corpus analysis tools for data assessment. Training has been part of their internship program at the research centre CODHUS (*Centre for Corpus Related Digital Approaches to Humanities*) at the West University of Timișoara, Romania, or it has been integrated into workshop modules offered in the framework of ongoing CODHUS research projects.

The corpora that have been used as a basis for linguistic analyses fall into several categories. Most of them are English-language corpora (BeYou, Kallo\$Sofia, COBG, ITOW, CHOCOP, CECS, HMC, VIGARECO), but there are also a few Romanian-

language corpora, according to the authors' interests (CEG, C-BLAGA, ADMIN). For more simple and efficient analyses, all the corpora are monolingual (only few corpora in Romanian contain an insignificant number of English terms). We noticed a high interest in oral or even multimodal corpora. Thus, one of the studies focuses on a diachronic perspective provided by photographs of the Harlem neighborhood of New York (Harlem), while three authors built their corpora focusing on vlogs and influencers' language (BeYou, Kallo\$\$ofia, CEG). In the latter case, the research involved listening, transcribing (with automatic or manual tools) and correcting the texts. It is widely known that, among the most difficult corpus analyses, are those related to audio archives. At the same time, the discourses contained in these three corpora are also influenced by strategies used to attract the public / views. Similarly, research on related cultural areas includes corpora such as HMC (consisting of movie scripts), COBG (containing lyrics of highly-rated songs), but also ITOW (Hollywood) or VIGARECO (gaming). Few authors have chosen to build literary corpora - C-BLAGA, CHOCOP - noting that C-BLAGA is limited to the poetic work of a single twentieth-century Romanian author, and CHOCOP relies on the diversity of authors and periods, "unified" by their main topic. The epistolary area is included in the CECS corpus, while the ADMIN corpus is the focus of an interdisciplinary study examining political discourse.

The topics and approaches covered by these micro-analyses demonstrate the dynamics of newer generations that no longer connect to sterile models and exclusively traditional methods, but are interested in current, genuine topics, innovation and, most importantly, analyse the contemporary world and cultural-literary phenomena with a critical perspective. In this sense, CODHUS does not remain an exclusivist center, but assumes its formative role and focuses mainly on applied research that can inspire the next collaborators – as we hope, as young and numerous as possible.

The corpus related quantitative and qualitative methods employed throughout the chapters are mainly supported by the open-source applications for linguistic research #Lancsbox, Voyant Tools and AntConc, and by the subscription-based tool Sketch Engine, made available through the CODHUS research centre. Most case-studies use word frequency lists, collocation patterns, and frequency distribution plots to generate meaningful frequencies of words and phrases that are visualized graphically using functions such as Word Cloud (Voyant Tools), GraphColl (#Lancsbox) or Word sketch (Sketch Engine). The quantitative results are complemented with in-depth exploration of relevant concordance lines extracted with the help of the Keyword in context (KWIC) function supported by all the tools. Lastly, some studies include sociological and historical approaches for the discussion of their findings, particularly the case studies using non-textual data (e.g. photographs) or a written biographical corpus.

The structure of the book

EBRIS

We know
books

The book is organized in two parts: Part I consists of chapters written in English and Part II consists of the Romanian version of the chapters. Twenty-six chapters have been selected and included in the book.

The first article, authored by Andreea DINCĂ, highlights that the linguistic patterns employed by beauty vloggers have two effects on viewers: they establish and maintain connections and boost the beauty gurus' reach. Alexandru FOITOȘ explored the materiality of Lucian Blaga's poetry by combining a quantitative and qualitative analyses. The author outlines the repeated material elements and their effects in shaping Blaga's literary expressions. While investigating the discourse particularities in interviews belonging to Romanian politicians and administrative personnel, Elena-Mirabela GRAMA identifies that the speakers' speeches comprise mostly of personal attacks and electoral promises, and not constructive criticism, fact which hinders inter-party collaborations. Adela-Diana CHINDRIȘ highlights the changes that occur in the Harlem of the 1930s to the 2010s, as depicted in photographs from different decades. The author revealed how Harlem's identity is re-shaped by both internal and external influences.

Roxana-Delia PROHAP analyses the language particularities of top Billboard hits from 1955 to 2019. The results indicate that the popularity of certain songs can be attributed to the use of recurrent grammatical and thematic features used by artists. Eliza BĂTRÎN continues the series of popular culture language and researches the linguistic features of British and American Hollywood actors. In this sense, the author identifies the dominant shared usage of certain discourse markers by both British and American actors. Florina-Maria ANDERCĂU identifies that influencers employ certain linguistic features to introduce beauty products to their audiences in a positive light, either explicitly or subversively. The results highlight that these linguistic particularities inspire their audience to use beauty products.

Margareta-Luisa BOICU analyses the use of language in disease poetry: the Plague, cholera, COVID-19. The conclusions display different perspectives of each illustrated topic, with darker depictions of the Plague and cholera. Teodora BOGOEVICI investigates the linguistic particularities of Elizabeth Cady Stanton's letters and her perspectives on women's rights. Digital tools help the author to identify Stanton's preoccupation towards women's rights and her reactions to misogynistic discourse. Bianca-Marinela SOCACI employs digital tools to discover the use of *gen* in multimodal texts (i.e. Romanian YouTube vlogs) and identifies that the selected word takes diverse hypostases: *gen* takes on the role of a filler word or means to highlight contrasts.

Karla-Claudia CSÜRÖS examines the linguistic characteristics of horror films and identifies that a viewer's attention span influences the complexity of the depicted horror atmosphere. Loredana BERCUCI illustrates a series of digital tools that can be used in Humanities research, with a focus on British and American studies. Digital Humanities and the applicability of the mentioned tools broadens the scope of previously-unexplored research perspectives. The volume concludes with Ciprian-Victor METEŞ's exploration of the thematic preferences of video game critics. The particularities of their perspectives reveal that critics especially examine the story and gameplay, with a focus on narratological elements. A series of specialised terms (technical ones) are also inferred in the textual representations.

Who needs corpus-related digital humanities research?

We hope that this collection of short research papers designed by students with the help of CODHUS teachers and researchers will aid humanities students and instructors in various humanities disciplines, such as literary studies, cultural studies, history, or media studies. Each of the projects provides an example of innovative approaches to humanities disciplines with methods borrowed from Digital Humanities that reflect the research interest of this generation of students. They can work as models for students who want to pursue research. Additionally, they could give instructors some insight into what topics draw students to research and provide an avenue for student-centered teaching.

In helping students design such projects, the CODHUS team hoped to develop a hands-on way for method acquisition. Furthermore, the skills developed by these independent projects can be integrated in different job market scenarios (e.g. journalism, social media analysis, IT tool development). This may broaden the students' choice of career after graduation in a field that is sometimes seen as providing limited tracks. Such a perspective can also increase the students' motivation to develop their skills when it comes to tools and methods related to corpus linguistics and Digital Humanities.

References:

- Brezina, V., & McEnery, T. (2020). Introduction to Corpus Linguistics. In *The Routledge Handbook of Second Language Acquisition and Corpora*, 11–22. London; New York: Routledge.
- Brügger, N. (2016). Digital Humanities in the 21st Century: Digital Material as a Driving Force. *DHQ: Digital Humanities Quarterly*, 10(3), 39–53.
- Curza, A. (2012). The electronic life of texts: Insights from corpus linguistics for all fields of English. In *English Corpus Linguistics: Crossing Paths* (pp. 7–21). Brill Rodopi.
- Dash, N. S. (2009). *Language corpora: past, present and future*. New Delhi: Mittal Publications.
- Golub, K., Kamal, A. M., & Vekselius, J. (2021). Knowledge organisation for digital humanities: An introduction. In *Information and Knowledge Organisation in Digital Humanities* (pp. 1–22). Routledge.
- Hinrichs, E., Hinrichs, M., Kübler, S., & Trippel, T. (2019). Language technology for digital humanities: introduction to the special issue. *Language Resources and Evaluation*, 53(4), 559–563.
- Johansson, S. (2008). Some aspects of the development of corpus linguistics in the 1970s and 1980s. In: Anke Lüdeling and Merja Kytö (eds.), *Corpus Linguistics. An International Handbook*, 33–53. Berlin: Mouton de Gruyter.
- McGillivray, B. & Tóth, G. M. (2020). *Applying Language Technology in Humanities Research: Design, Application, and the Underlying Logic*. Palgrave Macmillan.
- O’Keeffe, A., & McCarthy, M. (Eds.). (2010). *The Routledge Handbook of Corpus Linguistics*. Routledge.
- Reppen, R., Fitzmaurice, S. M., & Biber, D. (Eds.). (2002). *Using corpora to explore linguistic variation* (Vol. 9). John Benjamins Publishing.
- Roivainen, E. (2020). Generational changes in personality: the evidence from corpus linguistics. *Psychological reports*, 123(2), 325–340.
- Santos, D. (2019). Literature studies in Literateca: between digital humanities and corpus linguistics. *quot; In Martin Doerr; Øyvind Eide; Oddrun Grønvik; Bjørghild Kjelsvik (ed) Humanists and the digital toolbox: In honour of Christian-Emil Smith Ore Oslo: Novus Forlag 2019.*
- Tognini Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam: John Benjamins.
- Zottola, A. (2020). Corpus Linguistics and Digital Humanities. Intersecting Paths. A Case Study from Twitter. *América Crítica*, 4(2), 131–141.

TALKING THEIR WAY TO SUCCESS: HOW DO YOUTUBE BEAUTY VLOGGERS SPEAK?

Andreea Dincă

andreea.dinca@e-uvt.ro

1. Introduction: The Language of Beauty Vloggers

Have you ever noticed how frequently your favourite beauty guru says like and you guys? While watching such videos, I noticed this linguistic behaviour and came up with the idea to investigate the language of YouTube beauty vloggers. From the perspective of a language student, several questions came to my mind: Do famous vloggers follow a typical pattern when they build their discourse? Is there any connection between the words that beauty vloggers use and their success?



Why beauty vloggers?

I decided to focus on beauty vloggers because they act as influential role models for many of their viewers, which impacts their fans' consumer behaviour. In other words, if a vlogger positively reviews a particular product, there are more chances that their viewers will choose that product over several other ones. The power that beauty vloggers have over consumers has not been missed by cosmetic companies which often collaborate with beauty vloggers.

2. Method: How the Study Was Performed

As a language student, I was curious to investigate beauty vloggers from a linguistic perspective using Corpus Linguistics methodology. Therefore, an extensive database of transcribed YouTube videos, called a corpus, was created. As such, the corpus was then analyzed with the help of Corpus Linguistics software packages, like AntConc (Anthony, 2020) or WordSmith tools (Scott, 2020).

The corpus, called BeYou, comprises twelve viral YouTube beauty videos not older than five years. To make the corpus representative, I needed to consider criteria like popularity, English variety and gender of the vloggers. Following these criteria, I selected twelve most influential YouTube beauty gurus, six vloggers that spoke American English (three were female, and three were male) and six vloggers that spoke British English (of whom three were female, and three were male).

And like, like, like...

A fun fact about *like* is that it used to be associated with the *Valley Girl Speech* and seen as a marker of “superficiality and lack of intelligence” (Fuller, 2003, p. 369). Researchers also argue that it is a stereotype of American English, which is usually associated with the language of young people (Fuller, 2003, p. 369, 372). Now, however, this discourse marker is becoming more and more widespread among English speakers.

In the case of beauty vloggers, one might think that they have enough time to plan their speeches and then record their videos. It seems, however, that the discourse markers contribute to the interactive character of the video: the interaction is highly sought after by vloggers. They make their speech look informal and unplanned, taking the makeup videos a step closer to the traditional model of face-to-face interaction. In this way, vloggers look authentic and appeal more to their audience. In the simple *concordance* analysis (figure 1), we can observe the context in which *like* appears in the corpus.

Figure 1. Concordance output of discourse marker "like"

“Thank you”, “guys”, “let me know what you think”

Beauty vloggers address their audience in a friendly and familiar way, even if there could be an infinite number of potential participants at the reception end, who usually are not even familiar with one another. (Dynel, 2014, p. 37).

The most frequent term used by vloggers to address the viewers is **guys**. This word is usually accompanied by **you**, in constructions like **you guys**. These linguistic elements that include the audience into the product presentation process are called *engagement markers*. Vloggers use engagement markers to address their viewers and encourage their followers to interact with the content the beauty vloggers post, as shown in the example below:

“So I really really hope **you guys** enjoy this look and let me know in the comments what **you** think.” (Kathleen Lights)

A *Concordance Plot* type of analysis (figure 2) shows at which stage of the video vloggers reinforce their relationship with their viewers, with the help of the engagement marker *guys*: every rectangle stands for one YouTube video and

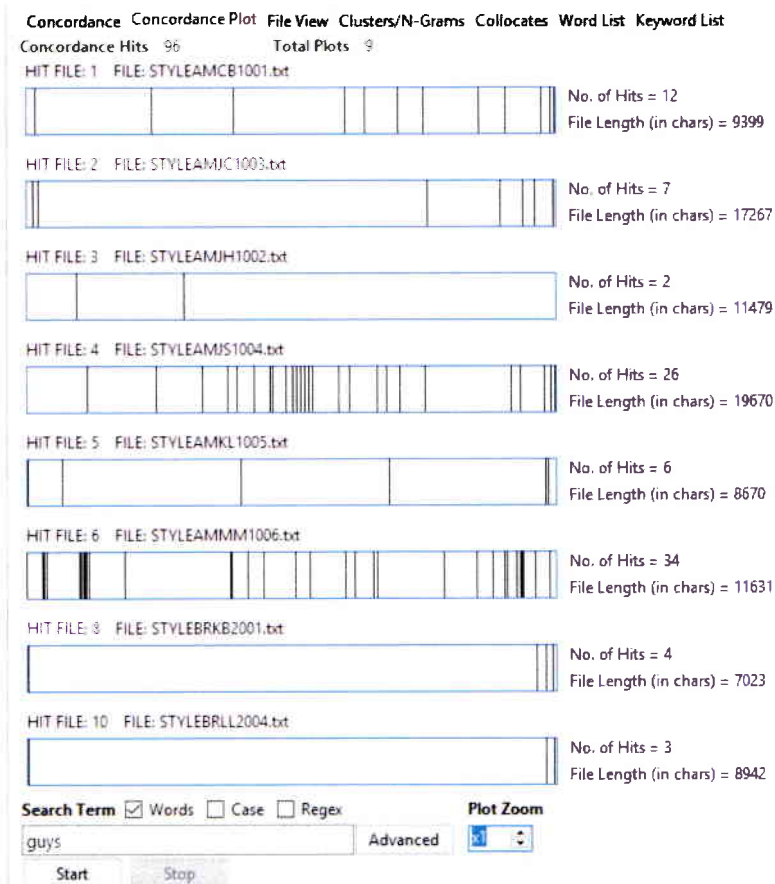


Figure 2. Concordance plot of engagement marker *guys*

every vertical line from inside the rectangle stands for one instance of the word “guys” and its position in the text. Some of the vloggers tend to use the word *guys* throughout the entire video, while others tend to use it at the beginning and end.

Another strategy that vloggers use to get close to their viewers is creating a common time and space for them and their audience, even if they might be thousands of kilometres apart. Aided by Youtube’s visual character, vloggers walk their viewers through all the steps of the make-up look as if the vloggers and their followers were in the same room. Words such as *this*, *these*, *them*, *here*, *now*, are frequent in the corpus, as illustrated in the following example:

“So I’m just going to use a little bit of Vichy Dermablend, this is number 25 (.) to cover this little scratch **here**.” (Lisa Eldridge)

These words are typical of face-to-face interaction because they would not have any meaning to the audience without the context provided in the video, just like in real-life interaction, where both the speaker and the listener share the same context.

Do male and female beauty gurus speak the same?

You might have seen, at least once, Jefree Star or James Charles rocking an impressive smokey eye makeup look. As you might have already noticed, YouTube has witnessed a massive rise in the number of male beauty vloggers along with men’s growing interest in makeup and cosmetic products. Even if putting on makeup is an activity traditionally linked to women, the male beauty vloggers promote the idea that makeup is gender-fluid and everyone should feel free to use it. The male beauty vloggers present themselves as a mixture of genders because they incorporate feminine and masculine features in the way they speak, act and dress. Even if they display feminine features, they do not deny their masculinity but rather celebrate it and fight for gender equality in the beauty world.

In my study, I looked at the language of male beauty vloggers and compared it to the language of female vloggers. Apart from the general features that all beauty vloggers seem to have, it turns out that the male beauty vloggers’ discourse differs from the one of female beauty vloggers.

The male vloggers’ fluidity of gender can be seen in how they mix both feminine and masculine words. For example, James Charles’s trademark greeting is “hi, sisters!” and he often calls himself “sister Charles”. In one video that I analyzed he calls himself “a fan girl” when talking about a favorite TV show. Manny Mua uses both masculine type words such as *man* (example 1) and more delicate ones like *pretty* and *glowy* (example 2).

1. “[...] Hey **man**, you should try this out [...]” (Manny Mua)
2. “[...] now looking at my skin, it looks **pretty**. It’s **glowy** [...]” (Manny Mua)

Another language feature that was discovered is that male beauty vloggers combine feminine words with coarse language. Words like *pretty*, *lightly*, *glowy* appear close to swearing words, like the *f-word*. My corpus analysis revealed a