ISBN: 978-84-1091-017-1 (PDF)
DOI: https://doi.org/10.14201/0AQ0373493500

# Thematic Corpus Construction, Representativeness, and Discursive Sustainability

## *Construcción de corpus temático, representatividad y sostenibilidad discursiva*

Jan BUTS
*University of Oslo*
*jan.buts@medisin.uio.no*

ABSTRACT: This paper discusses two corpora, the Genealogies of Knowledge Corpus and the Sustainability and Health Corpus. Both are constructed on the basis of topics and concepts, rather than factors such as genre or register. Selection criteria influence the kind of analyses one can apply to a set of language data, and the sort of conclusions one can draw from those analyses. Particularly relevant, in this respect, is the question of representativeness: how do we know that research results are meaningful beyond the data at hand? The contrast between ideal and pragmatic answers to this question is addressed in the paper's second section. The final part offers reflections on a recent complication. The production of text is increasingly relegated to automated systems. What does this mean for research principles based upon the assumption that the analysis of discourse can tell us something about the social world beyond the text?

KEYWORDS: corpus; concepts; representativeness; discourse analysis; automation; trust.

RESUMEN[1]: Este estudio analiza dos corpus que se formaron tomando como base temas y conceptos en vez de factores como género o registro. Los criterios de selección determinan el tipo de análisis que se puede aplicar a un grupo de datos lingüísticos, así como las conclusiones que se pueden desprender de dichos análisis. En este contexto, cobra relevancia la cuestión de representatividad: ¿cómo sabemos si los resultados de la investigación tienen sentido más allá de los datos que tenemos a nuestra disposición? El contraste entre respuestas ideales y pragmáticas a esta pregunta se abordará en la segunda sección del estudio. La última parte ofrece reflexiones sobre una complicación reciente.

[1] Thanks to Gustavo Góngora-Goloubintseff for the translation.

La producción de textos a través de sistemas automatizados va en aumento. ¿Qué significa esto para los principios de investigación basados en el supuesto de que el análisis del discurso pueda decirnos algo sobre el mundo social más allá del texto?

PALABRAS CLAVE: corpus; conceptos; representatividad; análisis del discurso; automatización; confianza.

## 1. THEMATIC CORPORA: WHY AND WHAT?

You can order a beer in a bar, but not in a court of law. You can file a complaint form at the hotel reception, but not in the family home. In any setting, norms and conventions regulate the scope of social interactions you can meaningfully engage in, and often we find that such conventions are codified as a set of practices operating in tandem with a set of linguistic expressions (Stubbs 2010). This is, in part, why the study of text corpora can tell us something about the functioning of social institutions. Parliamentary discourse, for instance, is constrained by, as well as constitutive of, the structure of acceptable decision-making processes within a particular political community. Similarly, the conventions of academic discourse hint at a common understanding of how to appropriately acquire and disseminate knowledge. In both the political and the scientific domain, then, we trust that we can study relevant texts and thus come to understand the state of affairs without directly witnessing or participating in any particular event. As long as we have access to a textual record, we can form an approximate idea of the proceedings that shape social reality. This does not mean, however, that society consists of a neatly compartmentalized assemblage of specialized institutions that transparently corresponds to a stable set of ossified scripts and statements. Our practices, and the language we use to describe them, are intermingled, open to negotiation, and in constant motion.

Consider, for instance, the concept of democracy. Arguably, we can identify a democratic political tradition that stretches back thousands of years, but both the practices and the attendant vocabulary that sustain this tradition remain «essentially contested», meaning that regular reinterpretation is an indispensable part of the concept's continued hold on the ideals and realities of political organization (Gallie 1956). Importantly, the scope and content of many foundational concepts, such as *democracy* or *citizenship* in the political sphere, or *evidence* and *expertise* in the scientific domain, are continually negotiated by a large variety of actors operating across a variety of settings without clear institutional boundaries. Studying the circulation and application of such concepts thus requires us to take into account a complex assortment of heterogeneous discursive environments. The Genealogies of Knowledge Corpus (GoK)[2] is a freely accessible language resource built in an attempt to do justice to this conceptual complexity. GoK contains material published in a variety of different formats, both online and offline, and places in conversation works produced across the centuries to ask, for instance, how «outsiders to the polity» can and could be represented (Baker 2020, 2). The corpus holds material in ancient Greek, medieval Arabic, Latin, and modern English, thus signalling

[2] https://genealogiesofknowledge.net.

that translation is a *sine qua non* for both the continuity and adaptability that simultaneously determine any concept's passage through space and time.

One could say, perhaps, that the study of concepts—contested or not—can benefit from the compilation of datasets that do not impose a strict discursive conformity on a layered textual realm characterised by multifaceted processes and sites of mediation (Baker et al. 2021). This is, then, one of the lines along which the conceptual research facilitated by GoK distances itself from more established practices in corpus-based translation studies: yes, one can study features of explicitation in a corpus of translated and edited alarm clock manuals published in 1995 by experienced professionals between the ages of 30 and 40, and then trace those features in a comparable dataset with a variable or two altered, and one is bound to find something of interest; lexical, pragmatic, perhaps even cognitive insights may be on offer. Yet, the closer one gets to complete factorial control, the further one distances the data from our everyday experience of language, the fragmented, teeming flow of semiotic stimuli arriving from God knows where directed at anyone who will listen. In other words, perhaps the study of a somewhat chaotic discursive reality requires somewhat chaotic corpora. Sealey and Pak (2018) describe, in this respect, the construction of a corpus aimed at examining the representation of non-human animals. Animals (wild, domesticated, feral, and figurative) are everywhere, leaping on and off the page, and so the researcher has to cast the net wide. Why not, then, collect and perhaps query, at the same time, newspaper articles on pests and interview transcripts on pets? There must be, in short, at least the possibility of thematic, rather than typological corpus work. This does not mean, however, that corpus design can proceed completely at random. Let us clarify with another example.

The Sustainability and Health Corpus (SHE)[3] is a language resource designed for both teaching and research, which expands the legacy of GoK, and is accessible via the same software environment. Whereas GoK is focused on constellations of concepts to do with the body politic and the language of scientific truth claims, SHE consists of texts broadly associated with the domain of medicine and healthcare. The corpus contains a range of works mostly published during the last half century: academic journal articles, blog posts, reports by non-governmental organizations, international treaties, books, and so on. The corpus grows gradually and organically, without a predetermined point of completion, but text selection does not happen haphazardly. Prioritised coverage areas, whose scope is frequently discussed among the corpus builders, include knowledge translation, pandemics and epidemics, health and environmental sustainability, sexual and reproductive health rights, as well as adolescent and young people's health. All texts are manually traced and annotated with contextually oriented metadata (e.g., publication date, organisation), which in turn aid data visualisation and the selection of smaller, more focused corpora for specific research or teaching purposes. SHE does not only store verbal information. Images found in the corpus texts are classified according to an encompassing taxonomy and made available via the concordance environment, so that the corpus facilitates the study of multimodality in a large variety of texts at the interface

[3] https://www.shecorpus.net.

of sustainability and healthcare. SHE is, in short, a carefully curated resource, but what does it ultimately represent? To what sort of external reality do we assume this collection of texts corresponds?

## 2. PATTERNING AND REPRESENTATIVENESS

In the grand scheme of things, corpus linguistics is still a relatively new approach to the study of language, and in seminal contributions written only a few decades ago, one encounters statements that are meant to reassure sceptics hesitant to entertain the possibility that theories of text should take into account empirical observations. Sinclair's work in particular, as cited below, illustrates that the study of corpora entails not just the application of an innovative method, but also the adoption of novel beliefs about what language is and does.

> The study of language is moving into a new era in which the exploitation of modern computers will be at the centre of progress. The machines can be harnessed in order to test our hypotheses, they can show us things that we may not already know and even things which **shake our faith** quite a bit in established models, and which may cause us to revise our ideas substantially. In all of this my plea is to **trust the text**. (Sinclair [1990] 2004, 23, emphasis added)

The plea entails that trust is placed in text at the initial expense of trust in one's own intuitions, mental models, or abstract theories. And one can be more specific about which aspect of «the text» deserves one's unreserved attention: corpus researchers are guided by «the search for—and belief in the importance of—recurring *patterns*» (Partington 1998, 9, emphasis in original). Patterns can be powerful guides. If we find, for instance, that any mention of *translation* in online news discourse is likely to be preceded by the words *lost in*, we learn something about the English language per se, but also about our cognitive mapping of abstract concepts, about news production, as well as, perhaps, something about public perceptions of translation practice (Buts and Malaymar 2023). Yet patterns are never immediately transparent. If we find that any mention of *fish* in conversations among friends is likely to be preceded by the words *plenty of*, we must remind ourselves that different realities and expectations are encoded in romantic and ecological discourse, and that providing hope is not always reconcilable with speaking the truth. Thus, text can be trusted, but only if subjected to a thorough background check. A context, in other words, must be established.

When we compile and study corpora, the patterns we encounter tend to be related to an identifiable «sample of a population of language users, a language variety, or a type of discourse» (Ädel 2020, 4). The types, varieties, and populations we consider meaningful say something about our conception of how the world works. One would be hard-pressed, for instance, to find a corpus of texts starting with the letter *t*, texts signed by people with a spotless complexion, or texts that from afar look like flies. Not every potential classification makes for a useful categorization, yet often the questions one aims to ask to determine which distinctions are meaningful: if we are lexicographers seeking to determine (approximately, of course) the thousand and one most frequent words in the Spanish language today, it does not matter whether the texts we consult were translated

from another language. Yet if we want to pose questions about the processes behind the consolidation of loanwords, it might. At this point, we reach the heart of the trouble with corpus design, namely the issue of representativeness. As explained by Biber:

> Representativeness refers to the extent to which a sample includes the full range of variability in a population. In corpus design, variability can be considered from situational and from linguistic perspectives, and both of these are important in determining representativeness. Thus a corpus design can be evaluated for the extent to which it includes: (1) the range of text types in a language, and (2) the range of linguistic distributions within a language. (Biber 1993, 243)

Ideally, for any language or variety thereof, we could estimate size, scope, and relevant characteristics, and apply statistical measures to gauge whether the chosen sample adequately represents the discursive set we aim to capture, in line with the questions we aim to pose. Biber (1993) offers a sophisticated procedure to this effect, yet some fourteen years after Biber's outline Leech (2007, 113) acknowledges that most corpus research is based on whatever «resources we have been able to lay our hands on». Not much has necessarily changed in this respect, as factors such as copyright concerns and limited resources often thwart noble scholarly aspirations. Yet, beyond worldly affairs, there is also the simple question of whether the game is worth the candle. As Halverson (1998) was quick to perceive, the idea of a representative corpus is fundamentally connected to a clear view of one's object of study. The downside of this insight: the «object of study» is often tied up with a particular research hypothesis, meaning that, ideally, for each question we have, we would need a separate corpus or, at least, a revision of relevant parameters, along lines whose silhouette we cannot accurately determine. All the while, we often use corpora to inform us not just about lexical patterning, but also about the whirling world at large. The discourse analyst is likely to relate semiotic action to social structure, but on what basis? What do we really know about the manner in which linguistic utterances impact how we live? Very little.

We do know that you can order a beer in a bar, but not in a court of law. Yet we do not necessarily know whether this statement becomes more or less convincing the second time you read it, and why. Just like we do not know which type of text is a more influential determinant of individual or collective behaviour: parliamentary debates, newspaper articles, or advertising campaigns. And we can scarcely prove, when we use resources such as GoK and SHE, that the discourse of *democracy*, *evidence*, or *sustainability* impacts the practices those concepts aim to shape. If it is already difficult to argue that a sample of texts about sustainability is representative of the entire set of relevant utterances, it is perhaps completely preposterous to claim that those texts are representative of something essential to sustainable development per se. And yet we trust the text. What compels us to do so?

## 3. NATURAL LANGUAGE AND SYNTHETIC MEDIA

A text, one could say, is a unit of «natural language used for communication» (Biber and Conrad 2009, 5). Text is produced under concrete conditions, at junctures of activity where people desire, in a broad sense, to cooperate towards a particular goal (Grice 1975).

In other words, language use has function and purpose, apparent in relation to a particular context of culture and context of situation (Halliday and Matthiessen 2014, 28, 32-3). This is, once again, why we assume we can trust the text to tell us something not just about itself, but also about human relationships and social institutions. Perhaps the case for representativeness is difficult to make, but the case for relevance is self-evident: people use language to intervene in their surroundings, and if communication was not effective, we would probably see less of it. The more so since language costs time and effort to produce. There is thus always, at least, a minimalist argument for corpus research: if enough data points towards a particular pattern, it is probably important and, one may venture, impactful in some way. It is always possible to argue against this position, but only in the contrarian fashion of the renegade archaeologist who asks: «Yes, the temple is the largest, most central, and most conspicuously adorned structure found on the site, but what if nobody ever went there?».

Concerns about corpus design are therefore secondary to the fundamental belief that texts contain some sort of noteworthy information, since humans have invested in the creation of communicative value. On this front, however, something is briskly changing. Large language models have made it possible to rapidly generate texts and images that mimic human communication. Synthetic media content, artificially produced and automatically disseminated, is already prevalent on social media, and may soon make up the bulk of online text production. Currently, much of the multilingual web already consists of machine-translated content. In the near future, the proliferation of machine-generated text will blur the relevance of any distinction between translation and other modes of content generation and adaptation. Furthermore, as operating systems and search engines increasingly rely on ad-hoc, personalized text generation, it is already hard to estimate the broader social relevance of highly customised communication flows (Buts 2021). Has anyone else ever read what you are reading?

Not to mention, in addition, our growing awareness that «the bulk of digital communications are no longer between people but between devices» (Woolley and Howard 2016, 4882). All of this is to suggest, of course, that perhaps we can no longer trust the text. Concern is not primarily with known issues such as inaccuracy or misinformation—humans never needed machines to lie—but with more basic communicative expectations. Do we still have an adequately stable notion of what natural language use consists of? And if not, how are we expected to study it? Are corpus researchers supposed to gather corpora composed of texts generated on the basis of language models trained on previously compiled corpora, and so on *ad infinitum*? Arguably not. Should students of language shun digital environments? No, the digital and the virtual have already become core drivers of cultural development and, thus, elements of nature. What we should do, however, is continuing to relentlessly interrogate the relationship between message and medium. If we believe in the study of patterns as a means of tracing meaning—no matter whether we aim to examine, say, grammatical features or thematic tendencies—we acknowledge, however hesitantly, that «there is ultimately no distinction between form and meaning» (Sinclair 1991, 7). This is the mystery that requires incessant pondering, particularly now that text has become a

throbbing, bulging assemblage, protruding through the crevices of our carefully constructed sense of existential propriety. As per the machine: it is important to note that.

## REFERENCES

Ädel, Annelie. 2020. «Corpus Compilation». In *A Practical Handbook of Corpus Linguistics*, edited by Magali Paquot, and Stefan Th. Gries, 3-24. Cham: Springer Nature Switzerland AG.

Baker, Mona. 2020. Rehumanizing the Migrant: The translated past as a resource for refashioning the contemporary discourse of the (radical) left. *Humanities and Social Sciences Communications* 6 (12): 1-16.

Baker, Mona, Jan Buts, Henry Jones, 赵文静, and 杨国胜. 2021. «用语料库考察概念的跨文化传播 - 知识谱系» 项目访谈». *外语教学与研究* 53 (1): 135-45.

Biber, Douglas. «Representativeness in Corpus Design». *Literary and Linguistic Computing* 8 (4): 243-57.

Biber, Douglas, and Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Buts, Jan. 2021. «Targeted Individuals: Personalised advertising and digital media translation». *Translation Spaces* 10 (2): 181-201.

Buts, Jan, and Deniz Malaymar. 2023. «A Look at What is Lost: Combining bibliographic and corpus data to study clichés of translation». *Corpus-Based Studies across Humanities* 1 (1): 1-22.

Gallie. W. B. 1956. «Essentially Contested Concepts». *Proceedings of the Aristotelian Society, New Series* 56: 167-98.

Grice, Herbert Paul. 1975. «Logic and Conversation». In *Syntax and Semantics, Volume 3: Speech Acts*, edited by Peter Cole, and Jerry L. Morgan, 41-58. Amsterdam: Elsevier.

Halliday, M. A. K., and Christian M. I. M. Matthiessen. 2014. *Halliday's Introduction to Functional Grammar*, 4th ed. London, New York: Routledge.

Halverson, Sandra. 1998. «Translation Studies and Representative Corpora: Establishing links between translation corpora, theoretical/descriptive categories and a conception of the object of study». *Meta* 43 (4): 1-22.

Leech, Geoffrey. 2007. «New Resources, or Just Better Old Ones?». In *Corpus Linguistics and the Web*, edited by Marianne Hundt, Nadja Nesselhauf, and Carolin Biewer, 133-49. Leiden: Brill.

Partington, Alan. 1998. *Patterns and Meanings: Using corpora for English language research and teaching.* Amsterdam, Philadelphia: John Benjamins.

Sealey, Alison, and Chris Pak. 2018. «First Catch your Corpus: Methodological challenges in constructing a thematic corpus». *Corpora* 13 (2): 229-54.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, John. 2004. «Trust the Text». In *Trust the Text: Language, corpus, and discourse*, edited by John Sinclair, and Ronald Carter, 9-23. London, New York: Routledge.

Stubbs, Michael. 2010. «Three Concepts of Keywords». In *Keyness in Texts*, edited by Marina Bondi, and Mike Scott, 21-42. Amsterdam, Philadelphia: John Benjamins.

Woolley, Samuel C., and Philip N. Howard. 2016. «Political Communication, Computational Propaganda, and Autonomous Agents: Introduction». *International Journal of Communication* 10: 4882-90.