Corpus Linguistics and the Social Sciences

Tony McEnery and Gavin Brookes

1. Introduction

The marriage of corpus linguistics and social science seems, initially, straightforward. Much work in corpus linguistics has oriented itself towards real world problems, including in areas such as climate change research (e.g. Dayrell and Urry 2015), criminology (e.g. Culpeper et al. 2017), defence studies (e.g. Germond et al. 2016), healthcare research (e.g. Bond et al. 2018), legal research (e.g. Lee and Mouritsen 2021), marketing (e.g. Fehrer et al. 2015) and policymaking (e.g. Mackney 2023). However, as with many areas where a superficial interaction can be evidenced by isolated studies, a true integration of corpus linguistics and the social sciences remains elusive. In part this relates to definitional problems – being clear about what we mean by social science and corpus linguistics may generate definitions that drive the two closer together or further apart. In a related point, our epistemology – that is, our theory of knowledge discovery – may also bring the two closer together or further apart, depending upon the position taken. Issues such as definition and epistemology, in turn, link to other major forces which militate in favour of or against integration; for example, data, tools and theory.

This paper explores these issues, looking, at an abstract level but illustrated through examples, at the interaction between corpus linguistics and the social sciences. We will begin by looking at epistemology, arguing that this is a major driver of corpus linguistics' integration with, or separation from, the social sciences. In doing so, we will outline our own epistemology and suggest a route that corpus linguistics may take in debates in the social sciences around epistemology. The route proposed should, in our view, maximise corpus linguistics' engagement with disciplines across the social sciences. However, we will also explore the varied nature of the social sciences, which is such that even a single discipline within the social sciences may exhibit significant internal variation in focus, theory and epistemology. We will see that such variation can militate for, or against, interaction with corpus linguistics. Throughout our discussion of epistemological concerns, we will note debates within corpus linguistics that echo these debates in the social sciences.

The paper then narrows in focus to look at a group of related areas in the social sciences that might have much to offer corpus linguistics. Following from that, we consider how data processing procedures in corpus linguistics – in terms of corpus mark-up, annotation and exploitation – appear to be converging, to an extent, with those in (especially qualitative) social science. In terms of quantitative analyses, we observe how there is much in common between corpus linguistics and the social sciences anyway. Likewise, we also consider how social science theory is exerting influence on studies in corpus linguistics. We conclude by reflecting on the nature of evidence, falsification and corroboration in corpus use in the social sciences.

2. Epistemological Fit

Since the mid nineteenth century at least, the idea of dealing with the social in the same way as the physical has been pursued (Comte 1858). This has given rise to an approach to the social sciences which can be placed under the broad banner of *naturalism*. This is the notion that the social may be examined in the same way, and by the same methods, as the natural world; or, to put it more succinctly, one may apply “the methods of physics to the social sciences” (Popper 2002: 2), in the belief that it is possible to derive laws which govern the social universe just as it may be possible to discern laws which govern the physical universe. Naturalism may be viewed as a form of *positivism*; an approach to epistemology based on “what is positively given avoiding all speculation” (Blackburn 2008).[1] Set against this is another broad approach to the social sciences which we may term *conventionalism*. This views the approach to the social via the methods of natural science, as being either mildly problematic through to being near impossible because of the nature of the object of inquiry. The social is subject to forces that the natural is not, such as free will. In other words, the object of inquiry is dynamic, not passive. Where social scientists place themselves between these two approaches, and within conventionalism in particular, is important for corpus linguists to understand. This is because this placement is a major factor in either promoting interaction with social science or militating against it. Some aspects of that interaction would be welcomed by corpus linguists, while others might not. This will become clear as we explore naturalism and conventionalism, and the distinctions between these, in more detail. That is the focus of the discussion presented in this section, which starts by considering naturalism.

There is, at least superficially, clear potential for alignment between corpus linguistics and naturalism. The approaches taken, for example, by researchers working in the field of social physics (see Jusup et al. (2020) for an overview) may appear familiar to corpus linguists, even if these are not necessarily used by all in the field. These approaches include Bayesian reasoning (Linka et al. 2022), network analysis (see Scott (2014) for an overview) and “big data” approaches (Ferreira et al. 2020). Yet the work in social physics tends to model reductively; that is, it simplifies social reality or limits the object of inquiry in order to produce results. Importantly, it tends to be normative, emphasising group behaviours over individuals, producing an approach to the social which may “oversimplify social dynamics and the diversity of individual characteristics that matter” (Kaufman, Diep and Kaufman 2020: 2) as “models used by physicists to describe social systems are too simplified to describe any real situation” (Castellano et al. 2009). In particular, as noted, the approaches are normative, averaging “over large societal groups [which] washes away individual peculiarities while retaining shared characteristics” (Kaufman, Diep and Kaufman 2020: 2). This has a profound impact upon our conception of what society is, as it becomes a collective and not the outcome of the interactions between individuals. While convenient for modelling, this is a conception of the social that we will not pursue here.

A further reason we will turn from the approach taken by naturalism is that the results have been, with regard to the study of language at least, often produced in the absence of input from linguists. As a consequence, such results have been, at times, naïve and insupportable from the perspective of linguistics. For example, a team of Danish and Japanese physicists analysed dialect maps developed in Japan to explore the diffusion of swearwords across Japan over time (Lizana et al. 2018). In this work, the dialect maps are used uncritically and in support of claims made about the diffusion of word forms across Japan. However, any

[1] Or, more accurately, epistemologically reductionist positivism – the ideas of the study of the social are being replaced with those of science to achieve a science-like social science.

linguist with experience of dialectology would have wanted the authors to consider the limitations of the type of data used – yet those limitations are not considered, and the data are barely presented in the paper, indeed it is treated almost as a given. The organization that developed the atlas data has written about its limitations, though, which include both the limited scale of the observations upon which the maps were based, and how it has led to the aggregation of data to a relatively high level in subsequent studies (Kumagai 2016: 334). A further limitation of this work, from a linguistic perspective, is that sociolinguistic features known to be important in language change, such as age and gender, go unremarked upon in the analysis. While naturalism has an initial appeal for corpus linguists, then, this appeal is ultimately undermined by naturalism's foundational approach, as this, in our view, mischaracterises society and often marginalises relevant expertise in favour of brute force modelling using tractable, rather than desirable, models.

Studies in social physics, such that by Lizana et al. (2018) described above, help to illuminate another distinction that is readily apparent within the social sciences; namely, the split between quantitative and qualitative approaches. This split should be better viewed as a continuum, as the two approaches are far from exclusive. However, naturalism, and its realizations (including social physics), is at the far end of the quantitative side of this divide, and quantification is one of the three features that define the approach to objectivity that such approaches take; specifically: i.) interest in the real; ii.) the claim to exclude values; and iii.) the use of methods oriented towards claims of absolute truth. The interest in the real we have discussed already; by limiting observations to the observable – the concrete – we can hold ourselves accountable to reality. The exclusion of values calls for a values-free approach to that data and to the explanations arising from the study of it – that is, whether something is right, wrong, desirable, undesirable, ideologically sound or unsound, should have no part in our reasoning. These features apply a natural science optic to the social and, in so doing, produce a distorted view of it. For example, while as a geologist I may well be able to physically observe a range of rocks, and plausibly argue that those rocks would exist whether humans existed or not, the same is clearly not true for social processes. Some of these are notional rather than concrete and are focussed upon the kinds of topics that have become the mainstay of research in corpus assisted discourse studies (CADS), such as identity, ideology and nationalism (Nartey and Mwinlaaru 2019: 220). Such concepts are, to a degree, subjective and difficult to categorize and measure. Such notional entities also have another property: they are bound to the subject as, unlike rocks, they would not exist without the observer, i.e. human beings, and they emanate from society, of which the observer is inextricably part. We will return to this entanglement of the observer and the observed in the social, shortly. A further point to note is that while a values free approach is an apparently reasonable goal, the call for a values-free approach may have consequences that we can all too easily value – we may get value-free conclusions that may prove harmful. Accordingly, at the very least, we should bring ethics as a set of values into research. For example, one might look back at a movement like eugenics and perhaps claim that it was simply driven by scientific observation and therefore it was value-free and was, as such, beyond critique. However, the social implications of the implementation of eugenics – racism, forced sterilization, bigotry and, ultimately, death camps – were not value-free and should have been considered through the optic of ethics earlier in the research cycle.

In the third feature of naturalism that we find another stumbling block for the corpus linguist engaging with the social sciences. The call for methods which permit a positivist approach is where quantification comes in as, regarding a values free approach it "is commonly believed that this is achieved in natural science by the use of quantitative methods, so social science should, as far as possible, follow the same path" (Montuschi 2015: 125). However, this approach partly negates the call for a value-free approach to study – the commitment of those taking an approach rooted in naturalism to an ideology of quantification is as marked a feature of their work as the commitment to ideological rectitude may be of a Marxist approach to society, for example. However, the values in question tend to be shared by policymakers, and it has been argued that there is an:

> "elite preference for a social science that is more scientific, positivist and analytical in its world-view. That this is an elite agenda in the sense of the expressed preferences of the leaders of peak associations is clear … (that) this is an elite agenda in the sense of the implicit preferences of the rich and powerful has often been suggested as well (Lather 2004). In this context the scientism of psychology and mathematicism of economics are seen as models to which the social sciences should aspire … more analytical social sciences that frame research questions in terms of mathematical symbols and answers them using quasi-experimental hypothesis tests." (Babones 2015: 455)

In such a context, it is hardly surprising that an association of quantification to positivism has been expressed and is well-attested in the literature, though it is not, as Babones later argues, an exclusive one for Sociology. We would argue the same as Babones (2015: 467) for the social sciences in general, including linguistics:

> "Sociology would benefit more from increasing the sophistication of its people than from increasing the sophistication of its statistics. Involvement in quantitative research is an important way for sociologists to increase their levels of sophistication, along with involvement in qualitative research, involvement in theorisation, involvement in teaching and involvement in public outreach. Interpretive and reflexive modes of engaging in all of these activities are more likely to result in the development of higher levels of sophistication and expertise than are positivist and unreflexive modes."

The forced "equivalence to quantitative criteria of 'good' research practice" (Mottier 2005) and "a common emphasis on attempts to formalize qualitative methods through the use of 'quasi-statistics' and software packages" (Mottier, ibid) bind computational and quantitative approaches, in the minds of many, to an epistemology rooted in naturalism. For corpus linguists, even if their orientation is more to the type of integration of quantitative and qualitative methods that Babones outlines, the link of quantification, one of the core features of corpus linguistics, to positivism is likely to be an important factor in limiting interaction between corpus linguists and social scientists. The best way to counter this is to shift to the epistemological middle ground. But to understand the need for that, one needs to understand the polar opposite of naturalism – conventionalism.

Late nineteenth century Germany provided the arena in which the social sciences split between positivism, as we have explored, and conventionalism.[2] Conventionalism embraces the intertwined nature of the natural and the social. As noted, conventionalism comes in a range of flavours, some of which may quite decidedly not be to the corpus linguist's taste – for example, while some philosophers in particular followed a route towards the study of the social based on naturalism, others reacted by moving to the opposite extreme. Positivism was rejected and replaced, by some, with the opposite belief – *antipositivism* (also called *interpretivism*). As antipositivists, conventionalists claimed that it was quite impossible to apply any of the methods of science to the study of the social. Indeed, the term positivism became "more of a term of abuse than a technical term" (Giddens 1977: 3) as a rancorous debate continued throughout the twentieth century. A significant problem for corpus linguists interacting with social scientists is that they are likely to encounter antipositivists and those antipopsitivists are likely to view corpus linguists as positivists, in part because of the orientation of corpus linguistics to quantification.

What is the principal antipositivist objection? It rests on an examination of one of the features of positivism already discussed – that positivists generally believed that the world existed independently of the observer and that through observation one could come to know that world. Through value-free, objective observations subsequently analyzed using the scientific method, positivists believed one could come to certain knowledge of the world. The critique of positivism runs along predictable lines – certainty in the social world is not a possibility, the social observer cannot be separated from the social object of inquiry, and this, in turn, gives rise to questions regarding the objectivity of the observations made. Quine (1961) makes two key arguments against positivism along these lines – that data is derived from senses and mediated by the concepts that we use to analyse it, and hence experience is subjective and our interpretations of it are subjective too. The observer and the observed are socially situated in space and time. The observer cannot step outside of social reality, of which they are part, to be objective. This drives away the possibility of data innocently presenting itself to the observer – the observer both perceives and interprets the data, and those perceptions and interpretations may, in turn, impact upon the conclusions drawn. This may then impact upon falsifiability. Antipopsitivists accordingly reject the scientific method, and empiricism, as an approach to the social sciences as a unified science that covers the social and physical sciences, as it:

> "fails because of the relationship between the social sciences and history, and the fact that they are based on a situation specific understanding of meaning … access to a symbolically prestructured reality cannot be gained by observation alone" (Outhwaite 1988: 22).

Corpus linguistics rarely considers epistemology as clearly as is common in many social sciences. This does not mean to say it does not have one – and when we look at debates in corpus linguistics, we may see the shadow of the positivist and antipositivist debates. A pressing example is the corpus-driven v. corpus-based debate. The initial statement of the approach has a clear footing in naturalism:

[2] For an account of this *Methodenstreit* and the role of Max Weber in proposing a middle way approach to the problems arising from it, see the discussion of Weber and economics in Maclachlan (2017).

> “in a corpus-driven approach the commitment of the linguist is to the integrity of the data as a whole, and descriptions aim to be comprehensive with respect to corpus evidence …. Theoretical statements are fully consistent with, and reflect directly, the evidence provided by the corpus …. recurrent patterns and frequency distributions are expected to form the basic evidence for linguistic categories” (Tognini-Bonelli 2001: 84)

This statement aligns well with positivism – there is no suggestion that subjectivity may have a role to play in forming our view of the data; rather, the data seems separate from the observer, allowing theory to “reflect directly” the evidence observed in an objective manner. Likewise, the process of interpretation of the data in ontological terms seems to be objective, with linguistic categories arising from distributional information. The rejection of that view was rooted in an acceptance of the impossibility of socially situated processes, such as language use, being viewed objectively, with the possibility of taking such an approach being dismissed by some corpus linguists as “an idealized extreme” (McEnery et al. 2005: 8) and its key assumption rightly dismissed because “the idea that empirical experience is the only guarantee of interesting theories was by and large abandoned long ago as a positivist error” (Stubbs 2013: 23).

While the terminology of the philosophy of science may be slightly alien to corpus linguists, then, the concepts are not. Moreover, it is advisable for corpus linguists to be cognisant of such concepts and the debates surrounding them, as these are likely to colour the perception that social scientists form of corpus linguistics. Indeed, in the literature we can see objections to corpus linguistics which follow the antipositivist critique, most notably with reference to context and subjectivity. To begin with context, it is unsurprising to see this critique, as it is precisely the critique which spawned the qualitative/quantitative divide in the first place (i.e., the belief that the social can only be studied in a broad, potentially amorphous context of which the observer is a part). It is context that forms the thrust of Baldry’s (2000: 36) critique of corpus analysis as “abstracting text from its context”. Decontextualization is a claim echoed by Thornbury (2010: 275), who claimed that corpus analysis gave only access to “surface features”. It is present, also, in Cameron’s (1998) claim that corpus linguistics examines too narrow a set of genres. Likewise, subjectivity arises in a range of forms in critiques of corpus linguistics – it is apparent in Widdowson’s (2000) claim that the disjunct between what is observed and what is, *a priori*, believed by the observer is a problem for corpus linguistics and, interestingly, in Borsley and Ingham’s (2002) criticism that it is the observer’s interpretations, not the data *per se*, which is the focus of linguistic analysis.[3]

[3] This last citation of particular interest to linguists as it shows a different approach to the criticism – Borsley and Ingham’s critique follows a Chomskyan tradition in its critique of corpus linguistics showing the potential affinity between a *scientia rationalis*, as opposed to the *scientia realis* approach taken by corpus linguistics, position and that of the antipositivists. The alignment is not complete, however – while the critique in favour of introspection denies the utility of corpus linguistics on similar grounds to those of the antipositivists, the *scientia rationalis* approach of the Chomskyans is clearly strongly aligned to a logical form of positivism. See Mcenery and Brezina (2022: 22-24) for a discussion of *scientia realis* and *scientia rationalis*. See Stubbs (2006, 2002) and Baker (2023) for a rebuttal of arguments against the use of corpora.

Concerns such as these can impact directly on the adoption of corpus linguistics in the social sciences. A good example of this is Zoldan (2024), who draws on arguments such as these, among others, when calling into doubt the utility of corpus analyses in the study of law and legal language, characterizing corpus approaches to the law as "this dream of objectivity" (Zoldan 2024: 403), with the problem with corpus linguists being "not their subjectivity, which may well be an inevitable part of the interpretative process, but the erroneous claim that they are superior because of their objectivity" (Zoldan 2024: 448).

So how a corpus linguist positions themself in the epistemological debate is important in interacting with the social sciences. Not all approaches to using corpus data stray towards positivism to the extent that the critique expressed by the likes of Zoldan holds. One such position, adopted here from McEnery and Brezina (2022), treads a path that many social scientists tread – namely, a line between the extremes of positivism and antipositivism. This approach, sometimes called *post-positivism*, treads a fine line between arguing for objective reality while also acknowledging the subjectivity that inevitably introduces the possibility of distortion into our observations of reality. McEnery and Brezina (2022) take a critical realist approach – they focus their work on reality (realism) but accept that only quasi-contact with it is possible because of the possibility of interference from the subjective (criticality):

> "critical realism, which acknowledges that we have quasi-contact with reality. As with scepticism, we allow the possibility of methodological relativism (or perspectivism), which recognises the complexity of reality and our imperfect grasp of the truth. This allows for multiple perspectives, or interpretations to compete in a rational debate when searching for the truth." (McEnery and Brezina 2022: 10)

Nonetheless, the best approach to studying linguistic reality is through the scientific method, an approach arising from the critical realism approach, giving rise to *critical rationalism*:

> with its emphasis upon argument and experience, with its device "I may be wrong and you may be right, and by an effort we may get nearer to the truth", is […] akin to the scientific attitude. It is bound up with the idea that everybody is liable to make mistakes, which may be found out by himself, or by others, or by himself with the assistance of the criticism of others. It therefore suggests the idea that nobody should be his own judge, and it suggests the idea of impartiality […] This is closely related to the idea of "scientific objectivity" […]. Its faith in reason is not only a faith in our own reason, but also in that of others (Popper 1945: 224-225).[4]

While this approach is only one of a number of possible approaches to explaining the epistemology of corpus linguistics, a clear approach to some of the questions outlined here is a good starting place for serious engagement with social scientists. Yet, the social sciences

[4] Note that linking critical realism to critical rationalism is distinct from Bhaskar's (1998, 2008) linking of critical realism to critical naturalism. That approach might be interesting for corpus linguists to follow, but in doing so they should be aware of critiques of this approach – see Zhang (2023) for example.

is a broad church. The extent to which different epistemological positions, and a focus on language itself, will engage social scientists should also be considered. This is the focus of the next section.

3. Disciplinary Scope and Interaction

It would be ideal to be able to align corpus linguistics and the social sciences and identify those areas which align and those which do not. However, such an approach would be naïve because the scope of the social sciences is uncertain. Also, within any subject within the social sciences, the orientation towards the study of language may vary. In addition, for any research area or even, perhaps, individual researcher, epistemological choices may differ. All of this complicates the formation of bridges between corpus linguistics and the social sciences.

The scope of the social sciences is both broad and indistinct. As such, deciding precisely what "counts" as a social science may be a vexing task. Inevitably, a degree of arbitrary choice is involved in deciding exactly which subjects to include under the label, and likewise in deciding which to exclude from it. However, no matter what parameters are used, the field is wide and varied. By way of example, the UK Economic and Social Research Council (ESRC) includes the following subjects in its definition of the social sciences: area and development studies; demography; economics; economic and social history; education; environmental planning; human geography; linguistics; management and business studies; politics and international studies; psychology; science and technology studies; social anthropology; social policy; social work; social statistics (including methods and computing); socio-legal studies; and sociology. Some of these subjects engage with the study of language to a degree (psychology and sociology, for example), some might conceivably have research questions to which linguists could contribute (social work and education, for example), while others may be focused so far away from language that the likely interaction with linguistics is marginal at best (social statistics, for example).

In one sub-field of corpus linguistics, CADS, alone we see substantial engagement with the social sciences, including with areas such as Business and Administration, Education, Health, Law, Politics and Religion (Nartey and Mwinlaaru 2019: 217). However, these labels represent aggregations of sometimes quite disparate subfields of study. Within any given area, the likelihood of interaction with linguistics can vary by sub-field. In psychology, for instance, developmental psychology includes language acquisition within its remit, an area where linguists in general and corpus linguists in particular may contribute. For example, in the UK a large team spanning linguistics and psychology work together in an ESRC research centre, the International Centre for Language and Communicative Development (LuCID),[5] exploring first language acquisition. In pursuit of their research questions, centre members draw on corpus-based findings (see Scholman et al. 2022). The integration of corpus data, linguistics and psychology is well established in this sub-area. On the other hand, the role of linguistics in industrial psychology is, at best, peripheral.

Sub-area may also impact not just on the question of engagement with corpus linguistics, it may even influence whether we view the subject as a whole as being part of the social sciences, or whether only parts of it truly are. For example, in linguistics, applied linguistics

[5] See https://www.lucid.ac.uk/ for more details about the centre.

is clearly a part of the social sciences while the study of phonetics and phonology is less clearly so. That is not to say that there is no role for phoneticians within the social sciences. Rather, the claim is that the scale and intensity of the interaction of linguistics with the social sciences varies as the subfields of linguistics and other social sciences interact. Thus, an engagement with the work of syntacticians and phoneticians, for example, is probably strongest in an area like developmental psychology and weak-to-non-existent across the rest of the social sciences. By contrast, the work of applied linguists, especially those working in discourse analysis, interacts quite strongly across the social sciences by comparison, with areas of particular intensity being education, management and legal studies, sociology and socio-legal studies. This has consequences for corpus linguistics – those areas which routinely draw upon corpus approaches, for example CADS (see Nartey and Mwinlaaru 2019, for an overview), the broad area of teaching and language corpora (e.g. Flowerdew and Brezina 2017) and corpus approaches to language and cognition (e.g. Gries and Stefanowitsch 2006; Lu et al. 2021) may find their work more broadly engaged with across the social sciences.

Importantly, the degree to which the engagement of social scientists with corpus linguistic research will occur varies, once more, according to epistemology. To focus once more upon discourse analysis, while work situated within critical discourse analysis, rooted largely in an anitpositivist tradition, is widespread across sociology, antipositivists undertaking critical discourse analyses are unlikely to accept the relevance of CADS. Indeed, within linguistics, Fairclough presents a defence of a critical discourse analysis which largely eschews corpus-based approaches (Fairclough 2015: 21-23) on largely antipositivist grounds, arguing instead for an epistemology closer to that of critical theorists such as Habermas. Others, by contrast, have critiqued Fairclough's approach precisely because of what is perceived to be an over-reliance on an antipositivist stance, which might allow practitioners to promote a political agenda based on cherry-picked examples (Widdowson 2009: 103-110). CADS, by contrast, has ploughed a furrow very much between the extremes of naturalism and positivism – it seeks analyses which accept linguistic reality, critically evaluates the data used, accepts that distortions may arise in analyses from a range of sources (e.g. subjectivity or issues of partial observation) and interrogates its corpus data in relation to the social context within which it was produced (Baker 2023). The epistemological divides that split the social sciences split linguistics too, then, and corpus linguistics' position relative to these splits will militate in favour of, or against, its engagement with individual social scientists, sub-fields of the social sciences, and the major subjects in the social sciences themselves.

The area of the social sciences that arguably comes closest to some of the methodological concerns of corpus linguistics is demographic studies and the related area of social surveys. The problems faced by such researchers are similar to those faced by corpus linguists – they often wish to characterise a population which is far too large to encompass fully. This leads to modelling of the population via sampling regimes. Demographers, for example, want to see how social and cultural factors impact upon that population. The variables are too many to define, leading to models of the population, and its interactions, being developed based on a sub-set of characteristics. Within the population, even with the limited set of characteristics observed, intersections between the characteristics give rise to further complexity. That complexity is, in turn, compounded because the interaction of the population with social and cultural variables also varies through space and time, leading to questions about how to measure change in the observed population over space (see Raymer et al. (2019), for example) and time (see Potrebny et al. (2017), for example) in a context where the variables themselves may change, either absolutely or by degree, over time (for example, a new disease appears and impacts on mortality, as happened with COVID-19).

If one were to conceive of language as one of the cultural and social variables that interacts with a population, then the link between the methodological concerns of demography and social surveys on the one hand, and corpus linguistics on the other, becomes clear. It is therefore surprising that interaction between these areas has been fleeting at best. This is in part because demographers in particular are not directly interested in language, but also because many linguists show minimal interest in some of the broader questions that demographers and survey-based social researchers ask, and hence do not necessarily see the value of their datasets. However, for corpus linguists, such datasets can provide some of the crucial context that would allow them to contextualise their observations. Some work in corpus linguistics has drawn on such resources, including panel survey data (e.g. Baker 2005; Blinder and Allen 2014)[6] and also demography, e.g. when using census data to design or assess corpora (see Love et al. 2017)[7] or using geo-demographic segmentation techniques to build a corpus sampling frame (see Crowdy 1993).[8] However, the engagement between corpus linguistics and such work is weak, even though the promise of approaching issues of social context are clearly held out by the resources and methods developed in such areas. They are one important way that corpus linguists can use triangulation (Baker and Egbert 2016) as a way of addressing context; that is, by situating corpus data, socially, in space and time. However, in that shift to triangulation, the methods of demography and social surveys, and the associated resources that have been built up around them in some countries, seem under-used.

Even where such data might be of use to a corpus linguist, the concerns of the demographer or social survey researcher are unlikely to be a direct point of contact for the corpus linguist. The demographer is typically interested in questions relating to patterns of births, deaths and marriages. These are approached largely through numeric data, often gathered either from public records or social surveys. While it might be conceivable that such information could be of importance to linguists – for example, those looking for cohort effects in language change might conceivably be interested in varying patterns of birth and death – most linguists would have no use for such data. Likewise, most social demographers have no use for data about language, except in so far as, perhaps, it represents a variable which might explain a feature of a social process they are examining, such as looking for alignments between inequality and language spoken by migrants (see Platt 2019: 135). Another exception is the role of narrative in social surveys. While not necessarily interested in the linguistic content of narratives, social surveys have, at times, focused on narratives told by subjects, which has

[6] Both used the British Social Attitudes Survey, a continuous national survey run since 1983. See https://natcen.ac.uk/british-social-attitudes

[7] See https://www.ons.gov.uk/census

[8] Crowdy used the ACORN, a classification which models and segments households in the UK. See https://acorn.caci.co.uk/

given narrative research a salience in social research and has generated data of potential interest to linguists (see Elliott 2005).[9]

So though apparently relatively distinct, at a level of abstraction, corpus linguistics faces similar challenges when constructing datasets to research in demography and social surveys. These fields may also help corpus linguists to gain quasi contact with the social context within which the language in a corpus is produced, helping corpus linguists to combat one of the (aforementioned) criticisms of their approach. Likewise, looking at the statistical processes run by demographers, social survey researchers and social statisticians, corpus linguistics may easily find familiar statistical procedures looking at familiar problems in familiar ways. For example, correlation, dispersion, distributions and regression are all statistical concepts and these, and related techniques, sit at the heart of much research focused on social surveys and demographic datasets (see Yusuf et al. (2014) for an overview of statistical techniques used in demography and Hanneman, Kposowa and Riddle (2016) for a more general overview of statistics in social research). For example, like corpus linguistics, observation in demography is closely linked to frequency and that, in turn, has introduced a Bayesian turn in demography (Bijak 2022), much as Bayesian processes are becoming more salient in corpus linguistics (see Stifter et al. 2022; Guajardo 2023; Woodin et al. 2024).

When looking to a subject like demography, it is possible that corpus linguists, seeing the scale of investment in the area, may assume that some problems that they have may be answered once and for all. For example, that demographers will have the answer to how to build a perfectly representative spoken corpus. However, a striking feature of demography research is that it is built upon the type of pragmatism that McEnery and Brezina (2022:67-68) called for:

> "There are certain properties that demographers would like their data to possess. Taken literally these desiderata are of the nature of ideals in that even in the most advanced countries they are never fully attained. Nonetheless, they are goals that should be kept I view". (Shryock et al. 1975: 4)

In many ways it is reassuring to know that an area that has attracted more investment than corpus research over a much longer period of time has not found the perfect solution to modelling a population either. Indeed, in demography as much as in corpus linguistics, models of a population are approximations. Again, this links back to epistemology. If our contact with reality, be that social and/or linguistic, is mediated through our datasets, then our contact is with quasi-reality (i.e., a proxy of that reality). This makes the extreme of naturalism less viable and suggests a post-positivist meeting place for corpus linguistics and the social sciences. However, in that meeting place, with similar challenges a fruitful interchange may occur, e.g. the methods of composing social survey panels, and especially panels which study the same cohort of subjects over time, are methods in the social sciences that corpus linguists should at least orient to and may very well benefit from.

[9] A further possible exception, in the UK, could be the inclusion of a question about the language spoken by the respondents. The move proved controversial, however, and far from being a point of engagement for linguists and demographers, it led to critique from linguists – see Sebba (2017), Wright and Brookes (2019) and Brookes and Wright (2020).

4. Data Processing

An obvious area where a fruitful cross-fertilization can occur between corpus linguistics and the social sciences relates to data processing and theory. There have almost been shadow developments occurring between corpus linguistics and the social sciences in this area. The most obvious area in which the social sciences have shadowed developments in corpus linguistics is corpus markup.

Corpus linguistics has long championed, and pioneered, markup schemes to permit the systematic encoding of metadata and interpretative analyses within corpora. Starting from a disparate range of mark-up schemes that were almost bespoke to individual corpus research centres (see McEnery and Wilson (2001: 34-38) for an overview of a range of early work on corpus markup schemes), corpus linguistics has kept pace with developments in textual markup to the extent that Extensible Markup Language (XML) is now almost a *de facto* standard in corpus construction. The precise set of entities to be used when creating a corpus still varies, from what might be called "maximalist" positions (such as the Text Encoding Initiative)[10] through to proposals for a so-called "core" set of elements for general use (see Caplan 2004) to an argument for a minimum range of elements that should be used when encoding a corpus (Hardie 2014). Likewise, some areas of the social sciences have also moved from bespoke markup schemes with data processing packages such as NVivo (Lumivero 2023) and Atlas.ti (2023) to XML. Meanwhile, other bespoke markup schemes, notably MacWhinney's (2000) CHAT markup scheme, have become capable of translation to XML. In CHAT's case that is achieved using the package CHATTER.[11] This means that, in principle, software packages which exploit that markup in the social sciences and those in corpus linguistics are more inter-operable than ever.

A key area where this interoperability is beginning to encourage cross-fertilization is in software packages used to annotate corpora. Nowadays, many concordance packages, such as SketchEngine and LancsBox, have annotation packages built in, enabling automated annotation of features such as parts-of-speech, parsing and even semantic annotation. Some packages help analysts introduce manual annotations, but these are fewer and less well developed, generally, that the automated systems. For example, in CQPweb it is possible to categorise examples according to an annotation scheme and then to use the scheme to explore the data. The emphasis of corpus software packages tends to be on quantitative exploration and automated annotation.

By contrast, in many ways software tools in the social sciences that facilitate textual analysis are oriented towards qualitative researchers and focus squarely on providing support for manual text annotation. Packages in the social sciences such as NVivo and Altas.ti are, arguably, very helpful packages for corpus linguists to use, especially where they are introducing manual annotations to corpus data, as this process is analogous to some of the typical uses of these packages (e.g. adding interpretative labels to interview data). The cross-fertilization of corpus linguistics and the social sciences via the use of packages like NVivo is now quite marked – at the time of writing, Google Scholar lists over 20,000 academic outputs that mention *corpus* and *NVivo*, covering research in areas as disparate as anthropology (Lukács 2021), business studies (Bengogo 2022), education (Matthews and Kotzee 2022),

[10] See https://tei-c.org/about/

[11] See https://www.talkbank.org/software/chatter.html

healthcare research (Greene and Brownstone 2023), social methods research (King 2010), social policy research (Jauffret-Roustide and Cailbault 2018), sociology (Sovacool et al. 2020) and tourism studies (Sanz-Blas and Buzova 2016). In some of these studies, the link between the two areas is fleeting – being simply the conceptualization of a dataset as a corpus (e.g. Bengogo 2022), for example. However, in others there is a much fuller attempt to carry out analyses which are clearly corpus-inspired using NVivo as a tool to undertake the work (e.g. Kotzee 2022, Sovacool et al. 2020). However, if we narrow our search of Google Scholar to focus more precisely on papers expressly acknowledging the influence of corpus linguistics by looking for *corpus linguistics* and *NVivo*, the results are still large and varied – counting over a thousand academic outputs covering a range of areas of the social sciences.

Nonetheless, what one can do with a corpus using a package such as NVivo is limited from the perspective of corpus linguistics, though, and this has consequences for studies in which only a package like this is used to analyse a corpus, of whatever size. NVivo has some strengths that corpus linguists should consider seriously. It is a good environment for adding annotations to a text, it has excellent multimedia capabilities, and it is XML-compatible. However, it also has limitations. Notably, NVivo is not a good source of frequency data, as the types of frequency lists that are common in corpus analysis packages are absent. Likewise, a host of techniques that many corpus linguists would want to use are absent, including collocation analysis and keyword analysis. This, of course, is not to say that NVivo is flawed; for what it was designed for – namely, coding in the context of relatively small-scale qualitative studies – NVivo is excellent. By the same token, where it exceeds the abilities of some corpus tools, it is not necessarily the case that those corpus tools are lacking; they were simply developed for a different set of users.

Used together, standard corpus tools and packages such as a NVivo could represent a powerful combination for users interested in building, manually annotating, and exploiting corpora. This is especially the case for multimedia corpora, as shown in Choubsaz et al. (2024), Clancy et al. (2023) and Shi and Khoo (2023), *inter alia*. In the past, exporting and importing data to those packages could prove difficult, because of issues of markup compatibility. However, with corpus linguistics and qualitative social science tools converging around XML, the possibility has opened up for the use of these programs to work together to support the relatively easy introduction and querying of annotations in corpus data. It is now easier than ever to exploit the overlapping and complementary needs of tools used by different research communities. With that said, there is little doubt that epistemology once again plays a role here. For example, as noted NVivo does not support well quantitative study because the users of NVivo have typically oriented to smaller scale qualitative analysis, at times on epistemological grounds. While a bridge has been built between tools like NVivo and corpus linguistics tools, then, we should not expect everyone to cross it.

While NVivo or other such tools can be used to input annotations, the question of what is annotated is more important than how that annotation is carried out. Put simply, this question goes something along the lines of, “what is our analytical scheme and what value does it have in aiding the process of interpretation?”. This is a question that is shared by linguistics and other areas of the social sciences, so it is perhaps understandable that annotation schemes is another area where there has been a flow of ideas between corpus linguistics and the social sciences. It is important to note that those schemes often arise from theory, so theories from the social sciences have become, at least indirectly, influential in corpus linguistics: corpus linguistics has provided insights into texts which have been interpreted through various theories, including cultural theory (Brookes and McEnery 2022), grounded theory (Curry and Péréz-Paredes 2023), poststructuralism (Brown 2024) and social network theory (Mackney 2023), for example. This engagement between corpus linguistics and a broad range of theory has been long predicted; McEnery and Wilson (2001: 193-194) argued that corpora could prove of use to a wide range of linguistic theories, for example. We are now seeing that, more broadly, as a method, or set of methods, corpus linguistics can be used within a broad range of theoretical frameworks both within and beyond linguistics.

There is a reverse flow from this trend, too. Each time the epistemological position of a researcher guides them to explore a theory through a corpus, the corpus plays a crucial role. More specifically, the corpus holds the theory accountable to quasi-reality and permits the possibility of falsification or corroboration (McEnery and Brezina 2022: 45-50).

Corpus users need, of course, to be mindful of what corroboration in particular means. It does not guarantee that the theory is "right", nor does it mean that this one theory alone will fit the data observed. The theoretical under-determination of corpus data ensures that this is the case (McEnery and Brezina 2002: 49). But there is an advantage to this again. The corpus does not select theories. Though it does reject some theories, it has the possibility of interacting with a range of theories across disciplines, where it can be used to test competing theories. As corpora develop over time, the cycle of attempts at falsification can continue. As this occurs, the corpus becomes an integral part of the research architecture of a discipline. In such a context, we should expect to see a wide range of theories guiding engagement with corpus data. If the theory touches upon questions of language, then the corpus is a key method, in principle, of seeking falsification or corroboration. The emergence of the corpus as a method in the social sciences is enabled by this conception of the corpus.

At this point, we also need to accept that in science as well as social science, the choice of theories which our data supports may be more sociological than scientific. Here, the ideas of philosophers of science such as Kuhn (1962) are more important to the ultimate adoption of a theory than the work of Popper, as "factual evidence is always insufficient to determine choice among scientific theories" (Cartwright and Montuschi 2015: 3), leading to the question of what determines theory choice when data has played its part. The answer to this may be varied:

> "Perhaps there are special virtues that all and only true theories can be expected to have, such as simplicity, coherence, and explanatory power. Or perhaps one theory is chosen over another because it has special advantages, maybe it solves problems that are particularly pressing at the moment, Maybe a worldview dictates. Maybe scientists get excited by the new ideas of the most recent theory or by the newest methods and concepts from other disciplines. Or perhaps adopting a particular theory serves some special interest groups over other or fits better with our view of prejudices" (Cartwright and Montuschi 2015: 3)

This is the context within which corpus linguistics finds itself in the social sciences; that is, more as filter than as final arbiter. Corpus linguistics permits theory that fits observations, but it does not determine uniquely which theory fits the observations. In that context, the use of corpus linguistics as a source of corroboration, in a context where social factors play in to the adoption and support of theory, is probably as important as its role in falsification.

5. Conclusion

The engagement between the social sciences and corpus linguistics is Protean – in some areas it is dynamic and growing, in others etiolated, in yet others non-existent and unlikely to grow. The varying linkage between the two can at times be explained simply by circumstance – the two areas have yet to gainfully interact, though they may in principle. However, there are also active barriers to interaction rooted in epistemology that are as intransigent and, in fairness, as principled as some of those that exist within linguistics which have stopped some linguists from using corpus methods. Over the years corpus linguistics has made progress in its home discipline by being results focused and by spawning new theories that use the data made available by corpus analyses. By showing positive results and new, productive theories, users of corpus data in linguistics have been able to tilt the scales in favour of corpus use. The same is possible in the social sciences.

However, in both linguistics and the social sciences other methods are also competing for attention. In particular data science, or "big data" methods, are producing results and demanding the attention of social scientists more broadly. The work produced by such researchers often aligns much more strongly with naturalism than corpus linguistics does. While this may, in light of what has been outlined in this paper, prove to be an opportunity for corpus linguists, allowing them to appeal to social science researchers who wish to engage with corpus data but not to shift towards positivism, it will only be so if two conditions are met, Firstly, corpus linguists need to be clear about their own epistemology. If they are not it is very easy to bracket corpus linguistics together with approaches to language data which, very often, are free of any serious reflection upon the nature of language in the social world. Secondly, corpus linguists need to be clear when marking this distinction. These two conditions have another, valuable, consequence. One thing that corpus linguists should be clear about – as should researchers using any method or set of methods – is that while a corpus can answer a range of questions worth asking, it cannot answer all questions that a researcher may reasonably have. Accordingly, corpus linguistics is bound to, and has been, used as one methodological approach amongst many in studies which use mixed methods and orient to triangulation. In being clear about what makes corpus linguistics distinct and what it has to offer, the role of corpus linguistics as one further methodological tool in the researcher's toolbox in the social sciences will be easier to make, and the engagement of corpus linguistics with the social sciences will be more easily facilitated.

6. References

ATLAS.ti. 2023. ATLAS.ti Mac Version 23.2.1 [Computer software]. Berlin: Scientific Software Development GmbH. Available from https://atlasti.com.

Baker, Paul. 2005. *Public Discourses of Gay Men*. London: Routledge.

Baker, Paul. 2023. *Using Corpora in Discourse Analysis*, 2nd edn. London: Bloomsbury.

Baker, Paul & Jesse Egbert (eds.). 2016. *Triangulating Methodological Approaches in Corpus-Linguistic Research*. London: Routledge.

Baldry, Anthony. 2000. *Multimodality and Multimediality in the Distance Learning Age*. Campobasso: Palladino.

Bengogo, Isidore Bimeme. 2022. Governance and organisational flexibility at the junction of African MFI’s sustainability issues. *Global Journal of Flexible Systems Management* 23(Suppl 1). S39–S50.

Bhaskar, Roy. 1998. *The Possibility of Naturalism: A Philosophical Critique of Contemporary Human Sciences*. London: Routledge.

Bhaskar, Roy. 2008. *A Realist Theory of Science*. London: Routledge.

Bijak, Jakub. 2022. *Towards Bayesian Model-Based Demography*. Cham: Springer.

Blackburn, Simon. 2008. *The Oxford Dictionary of Philosophy*. Oxford: Oxford University Press.

Blinder, Scott & William Allen. 2014. Constructing immigrants: portrayals of migrant groups in British newspapers 2010-2012. *Centre on Migration, Policy and Society Working Paper No. 117*. Oxford: University of Oxford.

Bond, Carmel, Gemma Stacey, Sarah Field-Richards, Patrick Callaghan, Philip Keeley, Joanne Lymn, Sarah Redsell & Helen Spiby. 2018. The concept of compassion within UK media generated discourse: A corpus-informed analysis. *Journal of Clinical Nursing* 27. 3081–3090.

Borsley, Robert D. & Richard Ingham. 2002. 'Grow your own linguistics'? On some applied linguists' views of the subject. *Lingua Franca* 112. 1–6.

Brookes, Gavin & Tony McEnery. 2022. Correlation, collocation and cohesion: A corpus-based critical analysis of violent jihadist discourse. *Discourse and Society* 31(4). 351–373.

Brookes, Gavin & Wright, David. 2020. From burden to threat: A diachronic study of language ideology and migrant representation in the British press. In: Paula Rautionaho, Arja Nurmi & Juhani Klemola (eds.), Corpora and the Changing Society: Studies in the Evolution of English, 113–140. Amsterdam/Philadelphia: Benjamins.

Brown, Katy. 2024. New opportunities for discourse studies: Combining discourse theory, critical discourse studies and corpus linguistics. *Journal of Language and Politics* (forthcoming).

Cameron, Deborah. 1998. Dreaming the dictionary: Keywords and corpus linguistics. *Keywords* 1. 35–46.

Caplan, Priscilla. 2004. *Metadata Fundamentals for All Librarians*. Chicago: American Library Association.

Cartwright, Nancy & Eleonora Montuschi (eds.). 2015. *The Philosophy of Social Science*. Oxford: Oxford University Press.

Castellano, Claudio, Santo Fortunato & Vittorio Loreto. 2009. *Statistical physics of social dynamics*. Reviews of Modern Physics 81.

Choubsaz, Yazdan, Alireza Jalilifar & Alex Boulton. 2024. *A longitudinal analysis of highly cited papers in four CALL journals*. ReCALL 36(1). 40–57.

Clancy, Cara, Emma McClaughlin & Fiona Cooke. 2023. Invisible animals: Exploring public discourses to understand the contemporary status of donkeys in Britain. *Anthrozoös* 36(6). 951–970.

Comte, Auguste. 1858. The Positive Philosophy of Auguste Comte. New York: Blanchard.

Crowdy, Steve. 1993. Spoken corpus design. *Literary and Linguistic Computing* 8(4). 259–265.

Culpeper, Jonathan, Paul Iganski & Abe Sweiry. 2017. Linguistic impoliteness and religiously aggravated hate crime in England and Wales. *Journal of Language, Aggression and Conflict* 5(1). 1–29.

Curry, Niall & Pascual Pérez-Paredes. 2023. Using corpus linguistics and grounded theory to explore EMI stakeholders' discourse. In Samantha Curle & Jack Pun (eds.), *Qualitative Research Methods in English Medium Instruction for Emerging Researchers*, 45–61. London: Routledge.

Dayrell, Carmen & John Urry. 2015. Mediating climate politics: The surprising case of Brazil. *European Journal of Social Theory* 18(3). 257–273.

Elliott, Jane. 2005. *Using Narrative in Social Research*. London: Sage.

Fairclough, Norman. 2015. *Language and Power*, 3rd edn. London: Routledge.

Fehrer, Julia, Sandra Smith & Roderick J. Brodie. 2015. Theorizing in Marketing Using Corpus Linguistics: A New Methodological Framework. In *Proceedings of the 44th European Marketing Academy Conference (EMAC)* [Online]. Available from https://www.researchgate.net/publication/289850414.

Ferreira, Paulo, Eder Pereira & Hernane Pereira. 2020. From Big Data to Econophysics and its use to explain complex phenomena. *Journal of Risk and Financial Management* 13(7). https://doi.org/10.3390/jrfm13070153.

Flowerdew, Lynne & Vaclav Brezina. 2017. *Learner Corpus Research: New Perspectives and Applications*. London: Bloomsbury.

Germond, Basil, Tony McEnery & Anna Marchi. 2016. The EU’s comprehensive approach as the dominant discourse: A corpus-linguistics analysis of the EU’s counter-piracy narrative. *European Foreign Affairs Review* 21(1). 137–156.

Greene, Amanda & Lisa Brownstone. 2023. "Just a place to keep track of myself": Eating disorders, social media, and the quantified self. *Feminist Media Studies* 23(2). 508–524.

Gries, Stefan Th. & Anatol Stefanowitsch (eds.). 2006. *Corpora in Cognitive Linguistics: Corpus-Based Approaches to Syntax and Lexis*. Berlin: Mouton de Gruyter.

Guajardo, Gustavo. 2023. Transitivity on a continuum: The transitivity index as a predictor of Spanish causatives. *Corpus Linguistics and Linguistic Theory* 19(2). 145–175.

Hardie, Andrew. 2014. Modest XML for corpora: Not a standard, but a suggestion. *ICAME Journal* 38. 73–103.

Jauffret-Roustide, Marie & Isabelle Cailbault. 2018. Drug consumption rooms: Comparing times, spaces, and actors in issues of social acceptability in French public debate. *International Journal of Drug Policy* 56. 208–217.

Jusup, Marko, Petter Holme, Kiyoshi Kanazawa, Misako Takayasu, Ivan Romic, Zhen Wang, Suncana Gecek, Tomislav Lipic, Boris Podobnik, Lin Wang, Wei Luo, Tin Klanjscek, Jingfang Fan, Stefano Boccaletti & Matjaz Perc. 2022. Social physics. *Physics Reports* 948. 1-148.

King, Andrew. 2010. Membership matters: Applying Membership Categorization Analysis (MCA) to qualitative data using Computer-Assisted Qualitative Data Analysis (CAQDAS) software. *International Journal of Social Research Methodology* 13(1). 1–16.

Kuhn, Thomas S. 1962. The Structure of Scientific Revolutions. Chicago: University of Chicago Press.

Kumagai, Yasuo. 2016. Developing the Linguistic Atlas of Japan Database and advancing analysis of geographical distributions of dialects. In Marie-Hélène Côté, Remco Knooihuizen & John Nerbonne (eds.), *The Future of Dialects: Selected Papers from Methods in Dialectology XV*. Berlin: Language Science Press. 333–361.

Lather, Patti. 2004. This is your father's paradigm: Government intrusion and the case of qualitative research in education. *Qualitative Inquiry* 10. 15–34.

Lee, Thomas & Stephen Mouritsen. 2021. The corpus and the critics. *The University of Chicago Law Review* 88(2). 275–366.

Linka, Kevin, Amelie Schäfer, Xuhui Meng, Zongren Zou, George EM Karniadakis & Ellen

Kuhl. 2022. Bayesian Physics Informed Neural Networks for real-world nonlinear dynamical systems. *Computer Methods in Applied Mechanics and Engineering* 402(9). 115346.

Love, Robbie, Claire Dembry, Andrew Hardie, Vaclav Brezina & Tony McEnery. 2017. The Spoken BNC2014: Designing and building a spoken corpus of everyday conversation. *International Journal of Corpus Linguistics* 22(3). 319–344.

Lukács, Gabriella. 2021. Internet memes as protest media in populist Hungary. *Visual Anthropology Review* 37(1). 52–76.

Lumivero. 2023. NVivo Version 14 [Computer software]. Available from www.lumivero.com.

Mackney, Sean. 2023. *Power and Discourse in the Policy Making Process*. Unpublished PhD Thesis. University of Bath.

Maclachlan, Fiona. 2017. Max Weber within the Methodenstreit. *Cambridge Journal of Economics* 41. 1161–1175.

MacWhinney, Brian. 2000. *The CHILDES Project: The database*. New Jersey: Laurence Erlbaum.

Matthews, Aiden & Ben Kotzee. 2022. Bundled or unbundled? A multi-text corpus-assisted discourse analysis of the relationship between teaching and research in UK universities. *British Educational Research Journal* 48(3). 578–597.

McEnery, Tony & Vaclav Brezina. 2022. *Fundamental Principles of Corpus Linguistics*. Cambridge: Cambridge University Press.

McEnery, Tony, Yukio Tono & Richard Xiao. 2005. *Corpus-Based Language Studies*. London: Routledge.

Mottier, Véronique. 2005. The interpretative turn: History, memory, and storage in qualitative research. *Forum: Qualitative Social Research* 6(2). https://doi.org/10.17169/fqs-6.2.456.

Nartey, Mark & Isaac N. Mwinlaaru. 2019. Towards a decade of synergizing corpus linguistics and critical discourse analysis: A meta-analysis. *Corpora* 14(2). 203–235.

Platt, Lucinda. 2019. *Understanding Inequalities*. Cambridge: Polity.

Popper, Karl. 1945. *The Open Society and its Enemies, Volume 2 – The High Tide of Prophecy: Hegel, Marx, and the Aftermath*. London: George Routledge and Sons.

Popper, Karl. 2002. *The Poverty of Historicism*. London: Routledge.

Potrebny, Thomas, Nora Wiium & Margrethe Moss-Iversen Lundegard. 2017. Temporal trends in adolescents' self-reported psychosomatic health complaints from 1980-2016: A systematic review and meta-analysis. PLoS ONE 12(11). [Online]. Available from https://doi.org/10.1371/journal.pone.0188374.

Quine, Willard V.O. 1961. *From a Logical Point of View*. New York: Harper & Row.

Raymer, James, Frans Willekens & Andrei Rogers. 2019. Spatial demography: A unifying core and agenda for further research. *Population, Space and Place* 25(4). [Online]. Available from https://doi.org/10.1002/psp.2225.

Sanz-Blas, Silvia & Daniela Buzova. 2016. Guided tour influence on cruise tourist experience in a port of call: An eWOM and questionnaire-based approach. *International Journal of Tourism Research* 18(6). 558–566.

Scholman, Merel, Liam Blything, Kate Cain, jet Hoek & Jacqueline Evers-Vermeul. 2022. Discourse rules: The effects of clause order principles on the reading process. *Language, Cognition and Neuroscience* 37(10). 1277–1291.

Scott, John. 2014. Social physics and social networks. In John Scott & Peter J. Carrington (eds.), *The SAGE Handbook of Social Network Analysis*. London: SAGE. 55–66.

Sebba, Mark. 2017. 'English as a foreign tongue': The 2011 census in England and the misunderstanding of multilingualism. *Journal of Language and Politics* 16(2). 264–284.

Shi, Jiayi & Zhaowei Khoo. 2023. Words for the hearts: A corpus study of metaphors in online depression communities. *Frontiers in Psychology* 14. 1227123. https://doi.org/10.3389/fpsyg.2023.1227123.

Shryock, Henry, Jacob S. Siegel, Charles B. Nam et al. 1975. *The Methods and Materials of Demography Volume I.* Washington: US Government Printing Office.

Sovacool, Benjamin K., Xiaojing Xu, Gerardo Z. De Rubens & Chien-Fei Chen. 2020. Social media and disasters: Human security, environmental racism, and crisis communication in Hurricane Irma response. *Environmental Sociology* 6(3). 291–306.

Stifter, D., Fangzhe Qiu, Marco A. Aquino-López, Bernhard Bauer, Elliott Lash, and Nora White. 2022. Strategies in tracing linguistic variation in a corpus of Old Irish texts (CorPH). *International Journal of Corpus Linguistics* 27(4). 529–553.

Stubbs, Michael. 2001. Texts, corpora, and problems of interpretation: A response to Widdowson. *Applied Linguistics* 22(2). 149–172.

Stubbs, Michael. 2002. On text and corpus analysis: A reply to Borsley and Ingham. *Lingua Franca* 112. 7–11.

Stubbs, Michael. 2013. Sequence and order: The neo-Firthian tradition of corpus semantics. In Hilde Hasselgård, Jarle Ebeling & Signe Oksefjell Ebeling (eds.), *Corpus Perspectives on Patterns of Lexis*. Amsterdam: John Benjamins. 13–34.

Thornbury, Scott. 2010. What can a corpus tell us about discourse? In Anne O'Keeffe & Michael McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics*. London: Routledge.

Widdowson, Henry G. 2004. *Text, Context, Pretext: Critical Issues in Discourse Analysis*. Oxford: Blackwell Publishing.

Woodin, Greg, Bodo Winter, Jeannette Littlemore, Marcus Perlman & Jack Grieve. 2024. Large-scale patterns of number use on spoken and written English. *Corpus Linguistics and Linguistic Theory* 20(1), 123–152.

Wright, David & Gavin Brookes. 2019. ‘This is England, speak English!’: a corpus-assisted critical study of language ideologies in the right-leaning British press. *Critical Discourse Studies* 16(1). 56–83.

Yusuf, Farhat, Jo M. Martins & David A. Swanson. 2014. *Methods of Demographic Analysis*. Cham: Springer.

Zhang, Tong. 2023. Critical realism: A critical evaluation. *Social Epistemology* 37(1). 15–29.

Zoldan, Evan C. 2024. Corpus linguistics and the dream of objectivity. *Seton Hall Law Review* 50(2). 401–448.