This is a pre-publication version of the following article:

Motschenbacher, Heiko. 2018. “Corpus linguistics in language and sexuality studies: Taking stock and future directions.” *Journal of Language and Sexuality* 7.2: 145-174.

When citing from or referring to this article, please use the final publisher version. This article documents work carried out within the LIDISNO (Linguistic Dimensions of Sexual Normativity) Project, which has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 740257.

Corpus linguistics in language and sexuality studies

Taking stock and looking ahead

Heiko Motschenbacher
Western Norway University of Applied Sciences, Bergen, Norway

As an introduction to the special issue, this paper presents an overview of previous corpus linguistic work in the field of language and sexuality and discusses the compatibility of corpus linguistic methodology with queer linguistics as a central theoretical approach in language and sexuality studies. The discussion is structured around five prototypical aspects of corpus linguistics that may be deemed problematic from a poststructuralist, queer linguistic perspective: quantification and associated notions of objectivity, reliance on linguistic forms and formal presence, concentration on highly frequent features, reliance on categories, and highlighting of differences. It is argued that none of these aspects rules out an application of corpus linguistic techniques within queer theoretically informed linguistic work per se and that it is rather the way these techniques are employed that can be seen as more or less compatible with queer linguistics. To complement the theoretical discussion, a collocation analysis of sexual descriptive adjectives in the Corpus of Contemporary American English

(COCA) is conducted in an attempt to address some of the issues raised. The concluding section makes suggestions for future research.

Keywords: corpus linguistics, queer linguistics, language and sexuality, methodology, sexual descriptive adjectives, collocation

1. Introduction

Corpus-assisted studies have recently started to enjoy a greater visibility in language and sexuality research, especially among European scholars. On the American side of the Atlantic, ethnographic approaches have played a predominant role in the formation and establishment of the field of language and sexuality since the 1990s. Of course, various other research traditions have additionally contributed to the field (for example, sociophonetics, lexicography, discourse analysis and applied linguistics), which corresponds to the multidimensionality that the relationship between language and sexuality exhibits. However, corpus linguistics, with its leaning towards quantitative methods, occupies a special position within the field as an important counterweight to work that is primarily qualitatively oriented.

Back in 1998, Cameron took a critical look at corpus linguistics that highlighted several of its (initial) shortcomings:

> Words, and more especially meanings, will always have a hidden history. While computerised corpora do make it easier to bring some aspects of that history to the surface (I think there is value, for example, in the collocational data they can provide), other equally important aspects may be more deeply buried as a result of the methods employed by the compilers and lexicographers: their sampling, their lemmatisation, their emphasis on the synchronic, even the sheer quantity of data they offer may be a hindrance to some kinds of analysis rather than a help. Perhaps the greatest problem implicit in the corpus dream, however, is its location within a powerful scientific or positivist discourse, whose own Keywords are “rigour”, “accuracy” and “objectivity”. (Cameron 1998: 45–46)

It is obvious that corpus linguistics has come a long way since then and that most of the problems raised by Cameron have in the meantime been widely recognised and satisfactorily addressed by corpus linguists. For example, nowadays a range of historical corpora is available for diachronic analyses. Similarly, many corpus linguists today acknowledge that their methodology is not fully objective or positivist and that it involves a significant amount of researcher input in the shape of motivated choices (for example, in terms of data selection, choice of search queries, corpus tools, significance measures and cut-off points, data interpretation and explanation, etc). However, what I would like to highlight in Cameron's quote is that it recognises that corpus linguistic methods are not a linguistic panacea but may actually be less well-equipped for certain types of analysis or language-related research topics. This is also true for corpus linguistics in language and sexuality studies more specifically. Without a doubt, corpus studies have substantially improved our understanding of the relationship between language and sexuality through frequency-based evidence, but at the same time there are certain aspects of the linguistic construction of sexuality that are less well-retrievable by means of corpus linguistic techniques, at least if the latter are used in traditional ways.

The present article hones in on the prominent role that corpus linguistics has started to play within language and sexuality studies since the turn of the century, recapitulates insights from earlier corpus linguistic research in this field and reflects on which directions corpus linguistic investigations of language and sexuality may take in the future. The following section provides an overview of the corpus linguistic work in language and sexuality studies that has been carried out mainly over the last 15 years. Against this background, Section 3 discusses the compatibility of queer linguistics as a central theoretical approach within language and sexuality studies with corpus linguistic methodology. Section 4 outlines the basic methodological issues for the empirical study presented in Section 5. The study conducts a collocation analysis of sexual descriptive adjectives in a major American English reference corpus, in an attempt to address some of the issues raised in the foregoing theoretical discussion. The concluding section (Section 6) makes suggestions for future research and presents a brief overview of the remaining articles in this special issue on Corpus Linguistics in Language and Sexuality Studies.

2. Taking stock: Corpus-based studies in the field of language and sexuality

When looking at the role that corpus linguistics (see, for example, Biber & Reppen 2015, McEnery & Hardie 2012, O'Keeffe & McCarthy 2012 for recent state-of-the-art overviews) has played within sociocultural linguistics, it becomes apparent that various subfields have not drawn on this methodology to the same extent. For example, while language and gender research has used corpus linguistics for sociolinguistic and discourse analytical investigations, the use of corpus linguistics in language and sexuality studies is largely limited to discourse analytical studies. One finds numerous corpus studies that focus on the discursive construction and representation of female and male people (e.g. Aull & West Brown 2013, Baker 2008, 2010, 2012, 2014a, 2015, Baker & Levon 2016, Carroll & Kowitz 1994, Kjellmer 1986, Motschenbacher 2013, Pearce 2008, Sigley & Holmes 2002, Taylor 2013) or on the way women and men use language (e.g. Barbieri 2007, Grimm 2008, Harrington 2008, Jiménez Catalán & Ojeda Alba 2008, Murphy 2010, Rayson, Leech & Hodges 1997, Schmid 2003, 2015). Sexuality-related corpus linguistic studies, by contrast, have concentrated on the discursive construction and representation of sexual identities, relationships, desires or practices via language, while sociolinguistic investigations of sexually defined social groups remain the exception (but see King 2009, 2015). Still it should be noted that there is no strong dividing line between these two types of investigation, as many studies incorporate both discursive and sociolinguistic aspects (with studies documenting language use in personal advertisements as a case in point).

There are several reasons for this situation. As far as practicality is concerned, it is much easier to integrate speaker or writer gender information in the mark-up of a corpus than information on the sexual identities or desires of language users. While gender is generally taken to be readily identifiable by the analyst or corpus compiler, the sexual identifications of language users are less obvious (at least if they are not explicated in the data by the subjects themselves), and it may be deemed ethically questionable to ask subjects to specify them. Another, more theoretically grounded reason is that the relationship between language and sexuality has for a long time been conceptualised in terms of discursive formation processes that take place across language users and usage contexts, rather than as a type of language use (or, more essentialised, a linguistic variety) associated with certain types of sexual subjects. In fact, notions such as gay male, lesbian or heterosexual speech varieties have been discarded by sociolinguists as too essentialist, as they ignore the immense variability and contextuality of the linguistic behaviour within these social groups (but see the British variety Polari as a plausible historical candidate; Baker 2002). Accordingly, the research questions that queer linguists ask today have changed from "How do certain sexually defined social groups use

language?" to "How is sexuality (including sexual identities, desires, practices and norms) discursively produced via language?"

In (critical) discourse analysis, corpus linguistics has been playing an increasingly prominent role (see Baker 2006, Mautner 2009, 2016), with researchers taking advantage of the benefits that this methodology provides: an incorporation of larger datasets (to overcome the criticism of cherry-picking telling examples), methodological triangulation with qualitative methods and multiple perspectives on the data, and reduction (though clearly not elimination) of researcher bias (Baker 2013: 180). It is, therefore, not surprising that corpus linguistic methods have also been used to explore a range of sexuality-related topics, mostly from a (critical) discourse analysis perspective. Such studies fall into three major thematic groups:

a) Studies focusing on the linguistic representation and communication of LGBT people:
   - analysis of gay men's online chat communication (King 2009, 2015)
   - analysis of the representation of gay men in newspapers, sexual health documentation and sitcoms (Baker 2005, 2014a)
   - analysis of the representation of trans people in the press (Baker 2014b)

b) Studies documenting public discourses associated with sexual relationships:
   - analysis of newspaper articles or parliamentary debates on the equalisation of the age of consent or same-sex marriage (Bachmann 2011, Baker 2004b, 2005, Findlay 2017, Love & Baker 2015, Vigo 2015)
   - analysis of marriage-related discourses as conveyed by the usage patterns of *bachelor* and *spinster* (Baker 2008)
   - analysis of communication in sex and relationship education (Sauntson & Sundaram 2016)

c) Studies concentrating on the role of linguistic practices in sexualised communication and communication about sexuality:
   - analysis of user profiles and personal ads on dating websites (Baker 2005, 2014a, Bogetic 2013, Leipold 2006, Milani 2013)
   - analysis of pornographic short stories and erotic narratives (Baker 2004a, 2005, Marko 2006, 2008, Morrish & Sauntson 2007, 2011, Wilson 2012)

- analysis of the use of love metaphors and verbs of love making (Archer, Culpeper & Rayson 2009, Manning 1997)
- analysis of the discursive construction of stalking communication, 17th century prostitution and sexual child abuse (Gales 2015, McEnery & Baker 2017, O'Keeffe & Breen 2007)

In short, previous corpus linguistic work on language and sexuality has overwhelmingly focused on the discursive construction of sexual identities (a), sexual relationships (b) and sexual desires (c).

3. Corpus linguistics and queer linguistic challenges

Language and sexuality can in principle be approached from a range of theoretical angles, not all of which are equally critical in their handling of the topic. Still, it is probably adequate to say that queer linguistics (e.g. Hall 2013, Leap 2015, Motschenbacher 2010, 2011), an explicitly critical approach, is today most prominent within language and sexuality studies, with individual researchers showing various degrees of explicit identification with its theoretical tenets.

Some of the theoretical developments that have affected language and sexuality studies in general have only marginally influenced corpus linguistic work on sexuality. Queer theoretical issues like the questioning or deconstruction of gender- and sexuality-related binarisms (e.g. Bing & Bergvall 1996, contributions in Zimman, Davis & Raclaw 2014), the linguistic de-essentialisation of gender- and sexuality-related categories or the foregrounding of non-normative aspects to weaken traditional, dominant discourses have been incorporated in much of the qualitative discourse analytic or ethnographically based work on language and sexuality. In (mainly) quantitative approaches such as corpus linguistics, their implementation appears to be more complex, as the following discussion aims to show. I will scrutinise five prototypical aspects of corpus linguistic methodology and discuss their compatibility with queerly oriented language and sexuality studies: its quantitative foundation, its reliance on forms and formal presence, its foregrounding of highly frequent features, its need for categorisation and its focus on difference.

3.1 Quantification and objectivity

Corpus linguistics is prototypically understood as a quantitative descriptive approach. Even though it is commonly sketched out as a methodology that potentially combines quantitative and qualitative procedures, it is evident that a quantification of forms is more central to it, also because it is generally carried out before more specific aspects are selected for closer qualitative inspection. A great advantage of corpus linguistic research is that it can cover large amounts of language data, thus facilitating a more comprehensive analysis that is not restricted to the (detailed qualitative) analysis of a limited number of examples and, as a consequence, provides results that are taken to be generalisable and replicable. Quantitative findings are in corpus linguistics regularly submitted to statistical significance tests, which may suggest a higher level of objectivity. In principle, corpus linguistics allows researchers to adopt a bottom-up approach to their data that is supposedly unaffected by preconceived theoretical notions. However, in practice such a strictly corpus-driven approach is hardly ever used, as most corpus linguists conduct their studies with certain aims or theoretical issues in mind that will inform their choice of data, tools, tool settings, analytical categories and interpretive strategies. This is, of course, equally true for queer linguists who draw on corpus methods. Their critical focus on the production of ideologies surfacing in the discursive construction of sexuality, and on associated normativities, discrimination, exclusion, stigmatisation and marginalisation, will usually lead them to adopt a problem-oriented corpus-based, rather than corpus-driven, approach (see also McEnery & Hardie 2012: 5–6).

The relationship between queer theoretically informed work and quantitative methods is a complex one, as the former would not normally grant such methods a higher level of objectivity, using them in triangulation with more qualitative types of investigation. Non-normative discursive practices, in particular, are often better retrievable via a rich contextualised analysis that takes a closer look at the local micro-context (as in ethnographic research) and/or the wider social, political and historical macro-context (as in discourse analytic studies). It could be argued that a quantitatively based generalisation of findings is of secondary importance in such a context, because the central aim is probably to document and exemplify non-normative aspects which are more limited in their occurrence but serve to challenge dominant discourses.

The central (maybe even only) instrument in corpus linguistics that can complement quantitative findings with qualitative kinds of investigation is concordance analysis. Concordances are compilations of all the tokens of a certain form in a corpus. Concordance

lines provide information on the immediate syntactic context in which a feature occurs. Viewed from an ethnographic perspective, this represents a relatively impoverished notion of context, but within critical discourse analysis concordance analysis may indeed make a useful contribution to linking quantitative findings to the wider social context via the detection of semantic preferences and discourse prosodies (see McEnery & Hardie 2012: 135–142).

3.2 Foregrounding of highly frequent items

As a methodology, corpus linguistics has considerably improved our understanding of how the usage frequency of linguistic features is involved in the discursive formation of sexuality. A concentration on highly frequent linguistic features is in general useful for the identification and critical analysis of those dominant discourses that are frequently overtly expressed. If we adopt a poststructuralist or Foucauldian notion of discourses as ways of seeing the world whose linguistic expression surfaces intertextually, there is a certain merit in showing that such discursive traces do not just occur in a single text or text passage but across a wider range of texts.

Since quantification generally serves as the entry point for corpus linguistic analyses, it is self-evident that high-frequency items are more likely to draw the researcher's attention, that is, what will be selected for further qualitative inspection will usually be something that has proven to occur relatively frequently in a given corpus. As researchers can only handle a certain amount of data, they will normally select cut-off points, aiming at "one-off or extremely rare word types being minimised" (Culpeper 2009: 36). This means that in the corpus-driven analysis of frequency lists only those forms that populate the top frequency ranks are incorporated, while low-frequency items are ignored, even though they may contribute cumulatively to a certain discursive effect or may represent traces of alternative discourses. This eradication of marginal linguistic phenomena can potentially be a problem in queer linguistics (see Baker & Egbert 2016: 195), which, in its deliberate attempt to adopt the perspective of the marginalised or peripheral, has an interest not just in majority but also minority patterns (see also Barrett 2014: 219). In queer linguistic corpus investigations there should, therefore, ideally be some space for the identification of less frequently or infrequently occurring patterns, which may be useful for highlighting alternative or silenced discourses or for challenging dominant discursive regimes.

But there is another way in which the role that frequency plays in corpus linguistics may pose a problem to language and sexuality studies. There is no neat correspondence between frequency of occurrence and the strength or entrenchedness of a certain discourse. Even though highly frequent linguistic features can often be plausibly linked to the formation of dominant discourses, it turns out that frequency does not directly translate into degree of entrenchment. This becomes evident when one views the relative strengths of competing discourses in relation to the frequencies of their linguistic traces.

In the Corpus of Contemporary American English (see Section 4 for more details), for example, one finds 10,543 tokens of forms that start with the letter sequence *homosexual** (the asterisk in the search query standing for any letter sequence of size zero or larger), while for *heterosexual** only 4,221 tokens can be found. A search for *bisexual** retrieves only 2,227 hits. Even though it is ultimately plausible to claim on the basis of these frequencies that homosexuality represents a more dominant discourse than bisexuality, analysts would probably be reluctant to conclude from this that homosexuality represents a dominant discourse vis-à-vis heterosexuality. This is the case because the dominance of heterosexuality is associated with a default status that, in effect, often causes heterosexuality to be not explicated but tacitly taken for granted, while same-sex sexualities as the marked cases are more likely to be explicitly oriented to in communication, maybe because they are deemed to be problematic (see Baker 2013: 188).

A similar point is made by Baker (2013) in a study in which he compared a corpus of abstracts from the Lavender Languages and Linguistics Conferences to a general American English reference corpus. Even though it can be assumed that a substantial share of the papers presented at this conference dealt with sexuality-related English language use in the US by white language users, the forms *English*, *America* and *white* did not turn out key in the keyword analysis (Baker 2013: 200–201), which means that a somewhat skewed picture of the aboutness of the conference papers emerges if frequency of occurrence is taken as the sole indicator. In other words, dominance at the discursive level can sometimes go together with lower frequencies of the formal reflexes of a certain discourse. Of course, this conundrum can to some extent be solved if we concentrate on more specific sexual discourses than homosexuality and heterosexuality as macro-discourses, but the issue is serious enough for queer corpus linguists to cultivate a certain degree of scepticism concerning the all-encompassing explanatory power of frequency and a routine of supplementing such quantitative with qualitative analyses (see also Baker 2016: 139).

3.3 Reliance on form and formal presence

The prominent role of frequency in corpus linguistics is connected to another characteristic of this methodology, namely its necessary reliance on form, formal identity and formal presence in corpora. In order to find features in a corpus, one has to use linguistic forms within search queries, and if one works with an untagged corpus, one has to rely entirely on forms. A prerequisite for meaningful corpus linguistic analyses is therefore formal sameness and stability. This fixation on linguistic forms may be problematic where functional variability or semantic change play a role – aspects that often receive greater attention in poststructuralist approaches such as queer linguistics.

A strong reliance on word-forms has the effect that corpus linguistics is particularly well equipped for identifying patterns at the lexical level (see Baker 2004a), while other linguistic levels (morphology, syntax, semantics, pragmatics) that may potentially also play a role in the formation of discourses are more difficult, though in most cases not impossible, to handle. Grammatical and semantic corpus annotation facilities (as for example implemented in the tool Wmatrix; Rayson 2008) have considerably facilitated the work of corpus linguists who are more interested in the analysis of grammatical functions or semantic aspects. Besides, analysts may use their own annotation schemes to spot and quantify pragmatic aspects that do not show a neat form-to-function mapping that could be exploited in search queries. This means that today there is hardly any linguistic aspect that cannot be investigated with corpus tools, even though the extent to which this can be done in an automatic fashion varies substantially.

These developments are also important for queer linguists, as sexuality-related linguistic features often concern the functional, semantic or pragmatic domains and cannot be fully retrieved by exclusively formally based search queries. However, there can be no doubt that implied meanings, contextual reference, semantic change, polysemy or non-literal language use represent special challenges to corpus linguistic inquiry, as they cannot be captured through a mere quantification of forms. The fact that harmful gender- and sexuality-related discourses such as sexism, heteronormativity, heterosexism and homophobia tend to surface in more subtle and less explicit ways today than in former times (see Love & Baker 2015, Mills 2008), therefore, also poses new challenges for the investigation of these discourses with corpus linguistic tools.

Corpus linguistics shows a strong leaning towards the analysis of individual lexical items and their collocates (see Mautner 2016: 157), while linguistic structures smaller or larger than the (orthographic) word-form are more difficult to grasp. In language and sexuality studies, this has had the effect that the discursive construction of sexuality has primarily been documented at the lexical level. However, a point can also be made for a more grammatical view of the linguistic representation of sexuality. For example, in connection with verbs of love making, a central point of interest is which participants are explicated and which syntactic functions they occupy in the various possible syntactic constructions (*he kissed her; she kissed him; they kissed; she was kissed*; see Ehrlich 2001 and, from a corpus linguistic perspective, Manning 1997). Such an analysis documents which social actors are perceived to possess more (sexual) agency and, connected to this, power. Historically speaking, the famous desire-identity shift in the conceptualisation of sexuality (Cameron & Kulick 2003) is also likely to have had grammatical consequences for the way we use language to communicate about sexuality. Robust empirical work that tests whether concomitant grammatical shifts have taken place is yet to be conducted. Still, one would probably expect to find a shift from a higher reliance on verbal constructions to express sexual matters (desire) to a greater use of adjectives and nouns as sexual identity labels (see also Motschenbacher 2014).

Finally, the necessity to rely on linguistic forms also means that corpus linguistics can de facto only analyse aspects that are formally present in a corpus, while formal absences and the question what could potentially have been used, but (maybe for strategic reasons) was not, are notoriously difficult to tackle with the use of corpus tools. But the "importance of what gets left out" (Kulick 2005) has long been recognised in critical discourse studies and queer linguistics (see also Love & Baker 2015, Partington 2014, Schröter & Taylor 2018). For example, certain grammatical constructions or lexical combinations that are in principle possible but do not or only infrequently occur in a data set may instantiate discourses that are perceived to be marked or non-normative. A corpus linguistic method that can be employed to make more detailed statements about formal absences and infrequent usage types is co-occurrence analysis, that is, the analysis of the occurrence of specific word combinations as central components of a larger semantic or conceptual field (see Motschenbacher forthcoming b).

3.4 Categories as analytical foundation

Quantification necessarily builds on categories that are established before the counting can take place. In corpus linguistics, such categories may involve language user groups, text genres, lemmas, parts of speech, semantic categories or other categories for which a corpus has been annotated. Categories are typically problematised in queer theoretically informed research (see Barrett 2014 for a queer linguistic critique of the discursive regimes governing formal linguistics), as they rest on the notion of intra-categorical homogeneity and thus cover up intra-categorical heterogeneity, prototypicality and normativity effects, and problematic category members, that is, phenomena that are of interest to queer linguists.

Even though category scepticism in queer-minded work potentially targets all categories, it is often socially relevant categories that form the centre of attention. In language and sexuality studies, gender- and sexuality-related binarisms (male – female; masculine – feminine; heterosexual – homosexual; gay – lesbian etc.; cf. Bing & Bergvall 1996) are the categories that have most intensively come under scrutiny, as they do not just facilitate harmful simplistic perceptions (such as "good vs. bad", or "primary vs. secondary") but also tend to support the marginalisation and stigmatisation of aspects that do not neatly fit into a binary scheme which normatively dictates opposition and incompatibility. The discursive predominance of binarisms has been challenged by linguistic evidence from non-Western cultures and earlier historical periods that highlight non-binary configurations (gender neutrality; more than two; category similarity and overlap; gender crossing; see contributions in Zimman, Davis & Raclaw 2014). However, Davis, Zimman and Raclaw (2014) note that binarisms should not (and probably cannot) be completely dropped from our analyses as explanatory tools. Instead they "advocate for a more complex and contextually grounded engagement with the binary" (Davis, Zimman & Raclaw 2014: 3).

Such a handling of binarisms may be more feasible in ethnographic approaches that observe local practices, but it is more difficult to achieve in a quantitative approach like corpus linguistics, in which the quantification is often based on the very binarisms that are supposed to be challenged. To address this problem, corpus linguists will have to find ways to question categories despite the fact that their research firmly builds on them. This can be achieved through a secondary qualitative analysis that highlights problematic category members and contextual meaning negotiation. Diachronic corpus analyses and comparisons of corpora with text material from different cultures and/or in different languages can be important, as it helps to foster an understanding for the historical and cultural specificity of categories. But maybe one can also think of more complex ways of analysis that avoid

viewing one's data through the coarse grid of the same binary categories that are actually meant to be questioned.

Another aspect of queer linguistic value that may be effectively studied using corpus tools is the demonstration that traditional gender- and sexuality-related binarisms (or categories more generally) may not just be locally less relevant or irrelevant (as many ethnographic studies have shown; see contributions in Zimman, Davis & Raclaw 2014), but may also lose some of their predominance more generally speaking. For example, various recent corpus studies have shown that lexically gendered forms are today less frequently used (see, for example, Taylor 2013: 97). This may be because such forms are dropped altogether (as is increasingly the case with courtesy titles like *Mr, Mrs, Miss, Ms*; cf. Baker 2010: 142–144), or because they are replaced with lexically gender-neutral alternatives (e.g. *policeman/-woman* being replaced by forms like *police officer* or *cop*; Baker 2010: 135). In a similar vein, Motschenbacher (2016: 20–22) found that the lyrics of Eurovision songs contained fewer lexically gendered nouns and pronouns than a general pop lyrics corpus, and that gender similarity rather than difference discourses manifested themselves in the distribution of such forms. The linguistic make-up of the song lyrics thus contributes to the effect of making the Eurovision Song Contest a less heteronormative context than pop music in general.

3.5 Difference focus

A final aspect about corpus linguistics that may be (and has been) viewed critically is its strong predisposition to identify differences, while similarities between linguistic data sets cannot be grasped equally well by comparative corpus studies and are, as a consequence, often ignored. This was traditionally not seen as problematic, as difference findings were considered more noteworthy or interesting, whereas studies that did not find substantial differences were, in general, not found to be worth of publication. Corpus linguistic methods are a safe bet in this respect, as the likelihood that one will find some sort of difference between two corpora is extremely high if not even one hundred percent. A popular corpus linguistic method that shows a strong orientation towards the detection of differences is keyword analysis (see Baker 2004a). It involves the comparison of two corpora at the lexical level and identifies word-forms that occur unusually (in)frequently in a given corpus when compared to a reference corpus (positive and negative keywords), thus overplaying differences between the two corpora.

More recent work in language, gender and sexuality has taken a critical stance on research that concentrates exclusively on the documentation of differences, arguing that similarity carries an enormous de-essentialising potential. In corpus linguistic analysis, similarity has for a long time been an elephant in the room, and it is only recently that corpus linguists have taken on the challenge to integrate similarity-targeting methods and concepts into their studies. In a seminal paper on similarity in corpus linguistics, Taylor (2013) points out that similarity regularly plays a role in the data selection process, since researchers aim at using comparative corpora that do not vastly differ from each other but only in one salient aspect, while other aspects (text genre, language user group, variety, time period etc) are kept constant across corpora in order to control potentially intervening variables. However, this similarity requirement may not be as important as commonly thought, as the evidence suggests that keyword analysis is quite robust even in the face of supposedly "bad" or small reference corpora (see Scott 2009). Baker (2011) introduced the notion of "lockword", that is, a word whose frequencies are similar across corpora. Another pertinent way to identify similarities between corpora can also be to not just compare two datasets but three or more (see Baker 2004a: 349). For example, instead of exclusively comparing two corpora of gay male and lesbian dating advertisements with each other, one could additionally compare them both to heterosexual advertisements and a general reference corpus (such as BNC or COCA). This procedure will not just show how the gay male and lesbian corpora differ from each other but also what they share.

It may be seen as a dilemma for those who take a more critical stance on binarisms and differences that comparative corpus linguistic methods invariably produce statistically significant differences, despite the fact that the similarities between two corpora may be overwhelming. So if corpus linguistics is a difference-oriented methodology, one needs to contemplate how such an orientation can fruitfully be employed in queer linguistic projects. Of course, an identification of differences is not problematic per se, and it can easily be reconciled with queer linguistic tenets if it is coupled with a critical analysis of the social effects of the discursive construction of such differences. Much of the corpus linguistic work that we have in language and sexuality studies today operates along these lines. However, there is an additional dimension of difference that may be more relevant to queer linguistic work and has so far played only a minor role in the field. Difference can also be highlighted within dominant gender- and sexuality-related categories, and can then be used to challenge hegemonic difference discourses and to show how dominant categories partially overlap (so that a strictly binary conceptualisation becomes suspect).

The way that differences are identified in corpus linguistics is commonly via inferential statistics (for example, for the calculation of keywords), that is, once a statistically significant frequency difference is found, it may be treated as a "real" difference. Queer linguists would probably show reservations towards such a mechanistic declaration of differences, echoing a criticism that has previously been voiced, namely that statistically significant differences need not automatically be culturally salient or socially recognised differences (see Cameron 1998: 41 on statistically based vs. culturally salient keywords). This begs the question whether we can actually claim that the statistically significant differences that we find between corpora are (large enough to be) also socially relevant. Or, viewed from a more ethnographic perspective: why dig up patterns that nobody found problematic in the first place? There is a danger of overrating newly detected statistically significant differences, or as Baker puts it "[w]e need to guard against viewing a statistical difference as an absolute or binary difference" (Baker 2012: 114). So maybe differences that can be identified using descriptive statistics (analysis of frequency lists and concordance lines) speak more to the notion of social relevance, as the analyst need not resort to the "magic" of inferential statistics to make differences relevant.

4. Methodological considerations

In the following, I present a corpus-based analysis of the usage patterns of sexual descriptive adjectives that are commonly used as identity labels. This is done with two major aims in mind: 1. to find out in how far these adjectives are actually used to write and talk about identities (rather than sexual desires, practices and other aspects), and 2. to analyse a range of forms and their usage similarities and differences, thus circumventing a binary research design. I use the Corpus of Contemporary American English (COCA) as a dataset for this study, which at the time of writing contains 520 million words of American English language use in the text categories spoken, fiction, popular magazines, newspapers and academic texts, and covers the years 1990 to 2015 (Davies 2009, 2010). The findings that this study presents need to be viewed as historically and culturally specific, as they are evidence for current sexuality-related discursive practices in the US and may not surface in the same way in other contexts. The study concentrates on the usage patterns of the eight most commonly used sexual descriptive adjectives: *queer*, *gay*, *lesbian*, *homosexual*, *same-sex*, *heterosexual*, *straight* and *bisexual*.

Corresponding to an earlier study on the nominal co-occurrences of these eight adjectives (Motschenbacher forthcoming b), the following six basic sexual usage categories are distinguished among the nominal collocates of the adjectives: identity, gender, partner, relationship, desire and practice. These categories were taken to cover central aspects of the discursive construction of sexuality and systematically integrated in the analysis. This has the benefit of allowing the analyst to make statements not just about collocational overrepresentation but also about less frequently occurring patterns and notable absences in the collocational data. In a more bottom-up fashion, I also allowed for additional categories to be treated as relevant in the analysis if they surfaced in the collocates of a particular adjective.

A collocation is a pair of words that often occur in each other's proximity (see McEnery & Hardie 2012: 122–133). Even though analysts may deduce collocations from an inspection of concordance lines, the today more common notion of collocation, which is also used here, draws on inferential statistics to decide which forms occur unusually frequently with a certain word. The collocation function in COCA was used to calculate nominal collocates one position to the right of the eight adjectives, relying on the tenet that the company a word keeps tells us something about its meaning potential and about how it is involved in the formation of discourses. The top thirty collocates of each adjective were used for closer inspection, to see how prominently the six conceptual categories outlined above are represented in the collocate lists and to check whether additional categories turn out to be relevant.

5. Usage patterns of sexual descriptive adjectives: A collocation analysis

Table 1 presents an overview of the nominal collocations of the eight adjectives (numbers after # indicate the collocate ranking):

Adjective	**Nominal collocates in the six major conceptual categories**	**Categories among remaining collocates**
Queer	**identity:** #6 child, #7 community, #12 youth, #19 Arabs, #24 Arab, #28 identity **gender:** #16 women, #29 boy	**politics:** #3 nation, #18 politics, #20 movement, #21 activists, #26 activism

	desire: #22 desire, #23 feeling **relationship:** #9 family ABSENT: partner, practice	**academia:** #2 theory, #4 studies, #11 theorists, #14 history, #25 theorist **art:** #5 accessories, #27 cinema, #30 art
Gay	**identity:** #4 people, #5 community, #15 person, #20 characters, #23 culture, #26 parents, #28 lifestyle, #30 youth **relationship:** #1 marriage, #7 couples, #12 marriages, #14 couple, #18 friends, #29 relationships **(male) gender:** #2 men, #6 man, #8 male, #21 males, #22 guy **practice:** #17 sex ABSENT: partner, desire	**politics:** #3 rights, #10 activists, #11 pride, #19 activist, #24 issues, #25 bashing, #27 liberation
Lesbian	**identity:** #2 community, #4 people, #8 youth, #13 students, #14 parents, #18 soldiers, #21 Americans, #23 identity, #24 groups, #25 Catholics **relationship:** #3 couples, #5 couple, #15 relationship, #16 families, #17 relationships **(female) gender:** #9 women, #11 mothers, #22 daughter, #27 mother	**politics:** #6 rights, #10 alliance, #26 activists, #28 feminist, #30 issues **academia:** #1 review, #12 studies

	partner: #19 lover **practice:** #20 sex ABSENT: desire	
homosexual	**practice:** #1 acts, #2 behavior, #4 conduct, #5 activity, #14 sex, #19 experience, #20 practices **identity:** #10 community, #11 persons, #13 lifestyle, #18 teacher, #24 identity, #26 teachers, #27 person **relationship:** #6 relations, #8 marriage, #9 couples, #15 relationships, #16 relationship, #23 unions **desire:** #12 orientation, #22 desire, #28 love, #29 inclination **(male) gender:** #3 men, #25 man ABSENT: partner	**politics:** #7 rights, #17 stigma, #21 agenda, #30 activists
same-sex	**relationship:** #1 marriage, #2 couples, #3 marriages, #4 unions, #5 relationships, #8 couple, #9 weddings, #12 friendships, #15 civil unions, #18 relationship, #20 friendship, #22 ceremonies, #23 married couples, #24 blessings, #26 wedding, #29 partnerships, #30 relations	

	desire: #6 attraction, #16 orientation, #17 desire, #19 attractions **practice:** #13 behavior, #21 experience, #25 touch, #28 experiences **partner:** #7 partners, #10 partner **identity:** #14 parents, #27 parent ABSENT: gender	
heterosexual	**relationship:** #3 couples, #4 marriage, #7 relationships, #13 married couples, #14 marriages, #16 relationship, #18 couple **identity:** #9 students, #20 youths, #26 peers, #28 parents, #29 community, #30 groups **gender:** #1 men, #2 women, #5 male, #17 males, #21 man, #25 woman **practice:** #6 sex, #8 intercourse, #12 contact, #15 activity, #27 behavior **desire:** #19 love, #22 orientation, #24 desire	**disease:** #11 transmission, #23 AIDS

	ABSENT: partner	
straight	**(male) gender:** #10 man, #15 men, #23 guy ABSENT: identity, partner, relationship, practice, desire	
bisexual	**identity:** #4 people, #5 adults, #6 youth, #7 youths, #8 teens, #10 populations, #11 individuals, #14 persons, #16 adolescents, #17 identity, #19 community, #21 black, #24 veterans, #27 respondents **gender:** #1 women, #2 men, #9 males, #12 male, #20 female, #23 females **practice:** #3 behavior, #18 experience, #28 behaviors, #29 contact **desire:** #13 orientation, #26 tendencies ABSENT: partner, relationship	

Table 1: Categories of nominal collocates directly following the eight adjectives in COCA

Six of the thirty collocates of *queer* belong to the category identity, which is the most prominent category for this adjective. Some additional categories turn out to be more central than the other five basic categories (gender, desire and relationship occur only marginally, partner and practice not at all). Five collocates belong to the semantic field of politics (*nation, politics, movement, activists, activism*); five others to the academic realm (*theory, studies, theorists, history, theorist*); three collocates can be grouped together under the category art (*accessories, cinema, art*). The centrality of these alternative categories shows that the usage

range of *queer* is atypical in that it extends well beyond the six basic categories into aspects that are maybe less immediately perceived to be connected to sexuality. In terms of notable absences, queerness does not seem to be conceptualised as a characteristic of sexual partners, and it is remarkable that it is also not conceptualised as a practice or something that is done, even though it is the only one of the eight adjectives that can also be used as a verb (*to queer something*). To summarise the most common patterns: queerness is mainly conceptualised as an identity, as the making politics in the name of this identity, and as a matter of academic discussion.

For the adjective *gay*, the category identity is most strongly represented (8 collocates), followed by relationship (6 collocates) and gender (5 collocates). Among the gendered collocates one finds exclusively male nouns, which indicates that lesbian sexualities are not really covered by the lexically gender-neutral form *gay*. In other words, *gay* has a strong male social gender bias. Another substantially represented category with 7 collocates is again politics (*rights, activists, pride, activist, issues, bashing, liberation*). The categories practice, partner and desire occur only once or not at all. The meaning potential of *gay* is thus both similar to and different from that *queer*, as it covers relationship and gender in addition to identity and politics. Judging from these findings, gayness is mainly conceptualised as something that people are, as something that can be politicised, as a sexual relationship type and as a particular type of masculinity.

The adjective *lesbian* has most collocates in the category identity (10 collocates), followed by relationship (5 collocates) and gender (4 collocates). It is apparent that within these three categories one finds a number of family-related terms (*parents, families, mothers, daughter, mother*), which indicates that the notion of the family plays an important role in the construction of lesbian identities and relationships. An additional important category is again that of politics (*rights, alliance, activists, feminist, issues*). The categories academia, partner and practice are only marginally represented, desire occurs not at all. Overall, the collocational profile of *lesbian* is similar to that of *gay* in that identity, relationship, gender and politics play prominent roles in both. Not surprisingly, the gender category shows exclusively female nouns for *lesbian*. Taken together, this means that lesbianness is mostly represented as an identity, as a field of political activity, as a type of sexual relationship and as a particular form of femininity.

The adjective *homosexual* exhibits a broad applicability range, with five categories being represented by four collocates or more. The most prominent categories are practice and identity (both 7 collocates), relationship (6 collocates), and desire (4 collocates). Sexual

politics again emerges as an additional category (*rights, stigma, agenda, activists*). This is the first adjective for which practice and desire play a central role. The category partner is not represented. In other words, *homosexual* covers almost the entire range of common sexuality-related concepts except for the domain partner. While identity was the most commonly used category for *queer*, *gay* and *lesbian*, *homosexual* is equally commonly associated with practices. In comparison to *gay* and *lesbian*, the relatively limited representation of gender is noteworthy. In the few cases where it co-occurs with gendered nouns, these are male, which suggests a socially male bias (similar as for *gay*).

For the term *same-sex*, the relationship category is highly dominant, with a total of 17 collocates. Desire and practice are also commonly represented, but clearly less frequently (4 collocates each). The categories partner and identity are only marginally represented. Gender is completely absent from the collocate list, and there are also no additional categories that emerge. In a way, *same-sex* is the form among the eight adjectives that is least likely to collocate with identity-related categories (i.e. identity, gender, partner, politics).

The adjective *heterosexual* has the highest concentration in the category relationship (7 collocates), but identity and gender are similarly important (6 collocates each). Practice is also well represented with 5 collocates. Interestingly, the category sexual disease emerges as a minor additional category with 2 collocates (*transmission, AIDS*). This is remarkable, because HIV infection is stereotypically rather connected to gay men than to heterosexual people, and suggests that heterosexual HIV infection is perceived as the marked case that needs to be explicated. It is also noteworthy that the gender category shows the highest number of collocates of all adjectives so far, which indicates that heterosexuality is more strongly connected to gender binarism than the other labels.

With respect to the categories of sexual relevance, the adjective *straight* exclusively shows male collocates from the gender category (*man, men, guy*). Other sexually relevant categories do not surface in the top 30 collocates of *straight*, probably because its non-sexual meanings (*a straight line, a straight answer* etc) are dominant overall. Still it is noteworthy that, judging from these collocational data, straightness in general seems to be a characteristic that is associated with men and not with women, that is, the adjective exhibits a similar male social gender bias as *gay* and *homosexual*. This in turn suggests that men are in general more likely to be represented in terms of their sexual orientation or sexual identity.

Finally, the term *bisexual* shows a strong leaning towards identity (14 collocates), and smaller focal points on gender (6 collocates) and practice (4 collocates). Desire is only marginally represented. The categories partner and relationship are absent from the collocate

list, which indicates that bisexuality is in general not perceived as connected to more stable sexual relationships (a characteristic that *bisexual* shares with *queer*). Within the identity category, it is evident that age seems to be a crucial factor in the discursive construction of bisexuality, as one finds various nouns denoting young people (*youth*, *youths*, *teens*, *adolescents*) and only one denoting adults (*adults*). The centrality of young social actors suggests that bisexuality is commonly conceptualised as a phase that young people may go through.

The overlapping common usage patterns of the eight adjectives are visualised in Figure 1, which shows categories represented by at least 4 collocates among the top 30 collocates of a given adjective (except for *straight*, for which gender is indicated as the only collocate category):

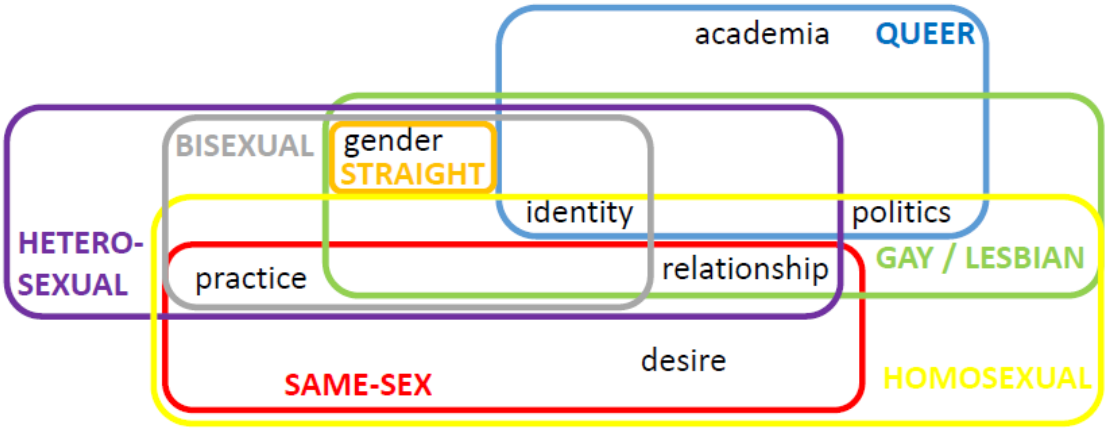

Figure 1: Conceptual map of sexual descriptive adjectives, based on collocation analysis of COCA data

Most significantly from a queer linguistic point of view, the picture that emerges in Figure 1 is anything but binary. It shows overlaps and specificities in the collocational usage patterns of the eight sexual descriptive adjectives, thus providing evidence of both similarities and differences alike. In terms of prototypicality, the categories identity, gender, relationship and practice turn out to be central aspects covered by (most) sexual descriptive adjectives, while desire, academia and politics are restricted to smaller subsets of adjectives. Somewhat surprisingly, the category sexual partner is not central to any of the adjectives and thus does not occur in Figure 1. Another remarkable fact is that the category desire only commonly occurs with two of the eight adjectives. This may be taken to suggest that, in contrast to earlier theoretical debates in language and sexuality studies (see Cameron & Kulick 2003),

desire plays a smaller role in the discursive construction of sexuality than the categories identity, gender, practice, relationship and politics. However, this would be a misleading interpretation, because sexual desire can be expressed by many other linguistic means apart from sexual descriptive adjectives, which have a stronger leaning towards the identity pole. The smaller role that desire plays in the nominal collocates of the eight adjectival labels, therefore, rather suggests that identity and desire conceptualisations may not go together that well.

Figures 2 and 3 present the same picture as Figure 1, but for smaller groups of adjectives. The three sexualities that at least partially involve female-male interaction (*heterosexual*, *straight*, *bisexual*) are presented in Figure 2, the remaining five sexualities in Figure 3.

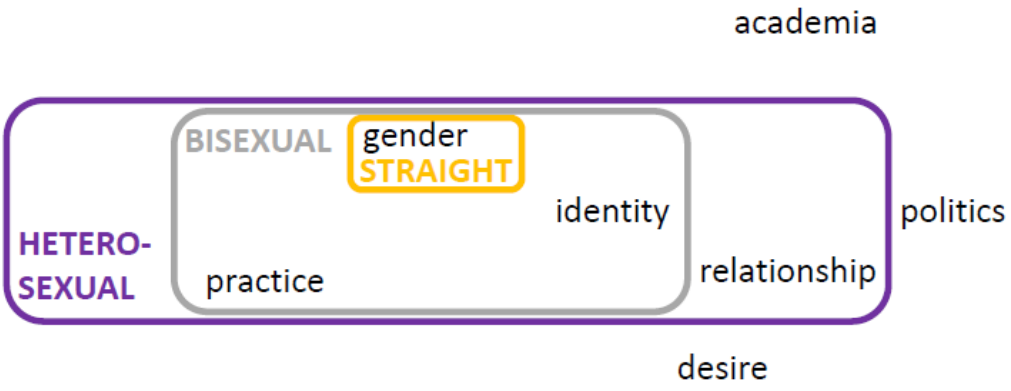

Figure 2: Conceptual map of *heterosexual*, *straight* and *bisexual*, based on collocation analysis of COCA data

Among the three adjectives in Figure 2, *heterosexual* shows the broadest range of application, spanning across the four prototypical categories identity, gender, relationship and practice. *Bisexual* is similar, but lacks an association with relationship. The adjective *straight* is exclusively associated with gender. Note that all three sexualities are connected to gender. In the case of *heterosexual* and *straight*, this could be taken as linguistic evidence for Butler's (1990) claim about gender binarism as a stabilising mechanism for the "heterosexual matrix." For *bisexual*, an association with gender binarism is also not surprising, as the morpheme *bi-* explicitly refers to such a binary. It is also noteworthy that, according to the collocation data, these three sexualities are apparently not deemed worth of explicit academic discussion or political activism. For *heterosexual* and *straight*, this probably has to do with the default status of these sexualities. For *bisexual*, this points to a marginalised status, with little

evidence of change. A similar point can be made about the absence of connections to the category desire. Maybe desire is by default taken to be heterosexual desire, so that such an explication is in most contexts not felt to be necessary. Bisexuality, on the other hand, rather seems to be denied the status of a (legitimate) desire (see also Thorne 2013).

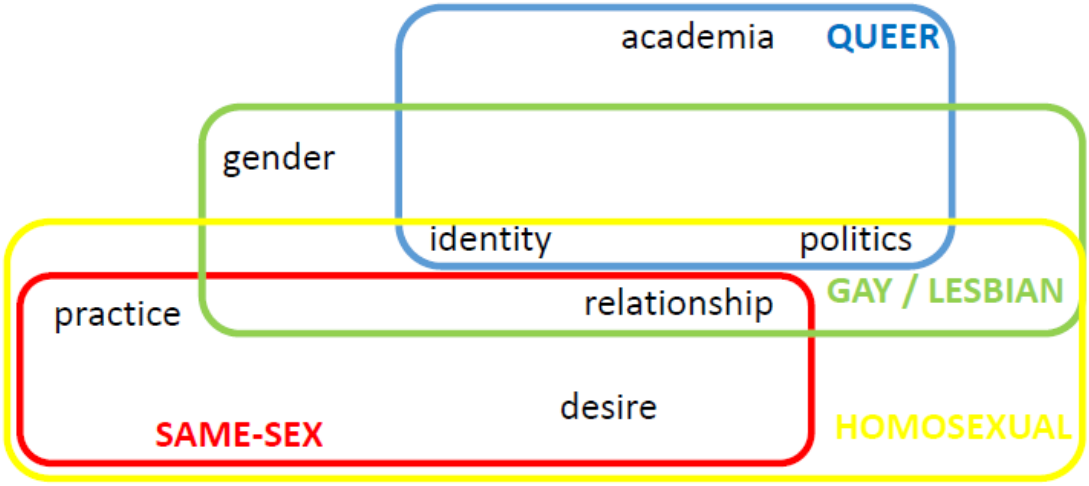

Figure 3: Conceptual map of *homosexual*, *gay*, *lesbian*, *same-sex* and *queer*, based on collocation analysis of COCA data

As can be seen in Figure 3, the five remaining non-heterosexual labels show a broader range of associations. The adjectives *gay* and *lesbian* exhibit identical application ranges (identity, gender, relationship and politics), with the gender category being female for *lesbian* and male for *gay*. Despite their asymmetrical genderisation (*lesbian* is lexically gendered; *gay* is lexically gender-neutral but strongly socially gendered), they seem to be used in a fairly parallel fashion along gender lines. Reference to sexual practices and desires is restricted to the adjectives *homosexual* and *same-sex.* The form *same-sex* shows the most idiosyncratic pattern in this group, because it does not cover gender, identity and politics, in contrast to the other four adjectives. What makes *queer* special is its additional connection with the academic realm. Note that three of the five adjectives in this group (*queer*, *homosexual*, *same-sex*) have no strong connection to gender.

The robust connection of *homosexual* to the realms of desire and practice is noteworthy, because the emergence of the term *homosexual(ity)* at the end of the 19th century is often interpreted as evidence for the historical desire-identity shift in the discursive construction of sexuality (see Cameron & Kulick 2003). Judging from the collocational

patterns of *homosexual* in COCA, the status of this adjective as an identity-denoting term becomes suspect. Maybe its associations with desire and practice are a more recent phenomenon – a question that future research should explore. Another reason may be that the use of *homosexual* as an identity label is still blocked to some extent by the pathological connotations of the word, which would explain why also other conceptualisations than identity play a prominent role in the usage patterns of this adjective.

6. Conclusion: Looking ahead

If we take the *queer* in *queer linguistics* seriously, the analytical focus of our research will be such that it enables us to take a critical view at the discursive mechanisms that contribute to the formation of sexuality and its normative regimes, including sexual identities, relationships, practices and desires. As the collocational analysis above has shown, a strong case can be made for identity-centered studies, but these need to be done in ways that do not further entrench dominant (binary) identity discourses. The documentation of identities that are (traditionally) considered non-normative plays a crucial role in this respect. Throwing a critical analytic light on how these identities are publicly represented via language is a legitimate and useful procedure, as it is likely to highlight aspects about the social perception of these identities that may be deemed ethically questionable (for example, a biased, negative, discriminating, stigmatising, stereotypical, incorrect, heteronormative, sexist or homophobic representation). However, as pointed out by Milani (2013: 617–618), a mere rhetoric of sexual tolerance or a promotion of the linguistic visibility of non-heterosexualities may not be sufficient for queer linguistic purposes. More thorough and ontologically oriented ways of resistance to the dominant sexual discursive regimes are necessary if queer linguistics is to distinguish itself from an LGBT-oriented linguistics. When we contemplate which kinds of corpus linguistic studies have the highest queer linguistic potential as far as de-essentialisation is concerned, it is probably studies that focus on language use in sexualised communication or communication about sexuality (see category c in Section 2). This is the case because they do not necessarily take identity categories like lesbian woman, gay man, heterosexual woman or heterosexual man as starting points.

As corpus linguistics is a methodology that traditionally shows a strong inclination to focus on quantitative description, linguistic forms and formal presence, high-frequency phenomena, categories and differences – in short, aspects that are more likely to provide

evidence for majority discourses – , its application in language and sexuality studies requires a high degree reflexivity (What are the discursive effects of methodological choices?) and creativity (Can corpus linguistic tools be used in ways that are conducive to rather than thwart queer linguistic aims?). An unreflected, traditional use of corpus linguistics is bound to possess only a limited de-stabilising and de-essentialising potential. More specifically, queer-oriented corpus linguists need to find ways to supplement and triangulate quantitative with qualitative modes of analysis (see Baker & Levon 2015, Baker et al. 2008, Marchi & Taylor 2009). They need to study the discursive construction of sexuality using non-binary and de-essentialising research designs. And they need to incorporate infrequent or formally absent phenomena indexing minoritised and silenced sexuality discourses in their analysis.

It is self-evident that corpus linguistics necessarily has to work with categories of some kind, otherwise quantification is not possible. The central question for queer linguists thus becomes which categories they can use as entry points for their research without sacrificing their theoretical convictions. If our main objective is to highlight damaging discourses in the public representation of LGBT or heterosexual subjects, using these identity categories as the basis for our studies can be a useful strategy. But when the goal of our studies is somewhat more on the de-essentialising or deconstructing queer linguistic side, it may be more useful to start our analysis from alternative sexuality-related categories, so that sexual identity categories are not automatically taken for granted or reinscribed. At the linguistic level, potentially relevant categories are, for example, verbs and verbal constructions that are used to express sexual practices or desires, body-part nouns and adjectives denoting physical features (and their eroticisation), or linguistic means of expressing sexual normativities (see also Motschenbacher 2014, 2018, forthcoming a). A stronger focus on text genre categories that are directly involved in the performance of desire (beyond dating sites and personal ads, genres like love letters, love poetry, valentine's cards, sex hotline talk, romantic literature and films come to mind) may also prove useful when it is our goal to de-essentialise identities.

The research design of the analysis carried out in this article was an attempt to approach sexual categories in a non-binary fashion, highlighting partial overlaps in the central conceptual categories sexual labels are associated with and systematically incorporating notable absences in the analysis. The detected usage patterns instantiate various sexuality-related discourses, ranging from global notions such as heterosexuality as default, over more specific conceptualisations such as gay or lesbian as identity or homosexuality as practice,

down to even more specific discourses like a family-orientation in the construction of lesbian sexualities or the role of age in the construction of bisexualities.

One limitation of the present study is that a collocation analysis can only cover word combinations that occur unusually frequently. This means that minority patterns and absences, that is, features that may play a role in the construction of marginalised and silenced non-normative sexuality discourses, could not be grasped in full. A co-occurrence analysis as performed in Motschenbacher (forthcoming b) seems to be better equipped for this purpose and may therefore fruitfully be employed in tandem with collocation analysis, to provide a richer picture of what can be found in the data. This picture can be further enhanced through an analysis of concordance lines, which yields insights into how certain collocations or co-occurrences are used (for example, in a positive or negative fashion). Another aspect that may be viewed as a limitation is that a very narrow window span was used in the collocation analysis. This procedure is legitimated by an effort to increase precision rates, that is, to make sure that a noun following a sexual descriptive adjective is in fact a head noun that is modified by the adjective. Future research may wish to explore whether the collocational patterns identified in the present study also hold for larger window spans. A final aspect that may be useful for queer linguistic analysis but is so far underexplored in corpus linguistics is the question whether there is something like the opposite of collocations, that is, words that repel each other or shun each other's co-presence. If such a notion could be operationalised, this would be another way in which absences could be more systematically integrated in the analysis.

The other articles in this special issue present and reflect on corpus linguistic analyses that explore the relationship between language and sexuality in novel ways. **Laura Paterson** and **Laura Coffey-Glover** triangulate analyses of keywords, semantic keyness and concordance lines to draw a multi-dimensional picture of the discourses connected to same-sex marriage manifesting themselves in a corpus of UK-based newspaper articles. **Lexi Webster** analyses discursive practices of self-sexualisation in a corpus of biography texts produced by various groups of gender-variant Twitter users. **Angela Zottola** focuses on the discursive construction of trans people in the UK press, using analyses of frequency lists, collocations and concordances to study representational practices. **Paul Baker's** contribution presents an overview of developments in corpus-assisted language and sexuality studies that highlights frequent misconceptions, potential areas of concern and future directions for the field.

Acknowledgement

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 740257.

References

Archer, Dawn, Culpeper, Jonathan & Rayson, Paul. 2009. Love – 'a familiar or a devil'? An exploration of key domains in Shakespeare's comedies and tragedies. In *What's in a Word-List? Investigating Word Frequency and Keyword Extraction*, Dawn Archer (ed), 137–157. Farnham: Ashgate.

Aull, Laura L. & West Brown, David. 2013. Fighting words. A corpus analysis of gender representations in sports reportage. *Corpora* 8(1): 27–52.

Bachmann, Ingo. 2011. Civil partnership – 'gay marriage in all but name'. A corpus-driven analysis of discourses of same-sex relationships in the UK Parliament. *Corpora* 6(1): 77–105.

Baker, Paul. 2002. *Polari - The Lost Language of Gay Men*. London: Routledge.

Baker, Paul. 2004a. Querying keywords: Questions of difference, frequency, and sense in keywords analysis. *Journal of English Linguistics* 32(4): 346–359.

Baker, Paul. 2004b. 'Unnatural acts': Discourses of homosexuality within the House of Lords debates on gay male law reform. *Journal of Sociolinguistics* 8(1): 88–106.

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

Baker, Paul. 2006. *Using Corpora in Discourse Analysis*. London: Continuum.

Baker, Paul. 2008. 'Eligible' bachelors and 'frustrated' spinsters. Corpus linguistics, gender and language. In *Gender and Language Research Methodologies*, Kate Harrington, Lia Litosseliti, Helen Sauntson & Jane Sunderland (eds), 73–84. Basingstoke: Palgrave Macmillan.

Baker, Paul. 2010. Will Ms ever be as frequent as Mr? A corpus-based comparison of gendered terms across four diachronic corpora of British English. *Gender and Language* 4(1): 125–149.

Baker, Paul. 2011. Discourse and gender. In *Continuum Companion to Discourse Analysis*, Ken Hyland & Brian Paltridge (eds), 199–212. London: Continuum.

Baker, Paul. 2012. Corpora and gender studies. In *Corpus Applications in Applied Linguistics*, Ken Hyland, Chau Meng Huat & Michael Handford (eds), 100–116. London: Continuum.

Baker, Paul. 2013. From gay language to normative discourse. A diachronic corpus analysis of Lavender Linguistics conference abstracts 1994–2012. *Journal of Language and Sexuality* 2(2): 179–205.

Baker, Paul. 2014a. *Using Corpora to Analyze Gender*. London: Bloomsbury.

Baker, Paul. 2014b. 'Bad wigs and screaming mimis': Using corpus-assisted techniques to carry out critical discourse analysis of the representation of trans people in the British press. In *Contemporary Critical Discourse Studies*, Christopher Hart & Piotr Cap (eds), 211–235. London: Bloomsbury.

Baker, Paul. 2015. Two hundred years of the American man. In *Language and Masculinities: Performances, Intersections, Dislocations*, Tommaso M. Milani (ed), 34–52. New York: Routledge.

Baker, Paul. 2016. Gendered discourses. In *Triangulating Methodological Approaches in Corpus Linguistic Research*, Paul Baker & Jesse Egbert (eds), 138–151. London: Routledge.

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

Baker, Paul, Gabrielatos, Costas, KhosraviNik, Majid, Krzyżanowski, Michał, McEnery, Tony & Wodak, Ruth. 2008. A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society* 19(3): 273–306.

Baker, Paul & Levon, Erez. 2015. Picking the right cherries? A comparison of corpus-based and qualitative analyses of news articles about masculinity. *Discourse & Communication* 9(2): 221–236.

Baker, Paul & Levon, Erez. 2016. 'That's what I call a man': Representations of racialised and classed masculinities in the UK print media. *Gender and Language* 10(1): 106–139.

Barbieri, Federica. 2007. Older men and younger women. A corpus-based study of quotative use in American English. *English World-Wide* 28(1): 23–45.

Barrett, Rusty. 2014. The emergence of the unmarked: Queer theory, language ideology, and formal linguistics. In *Queer Excursions: Retheorizing Binaries in Language, Gender,*

and Sexuality, Lal Zimman, Jenny L. Davis & Joshua Raclaw (eds), 195–224. Oxford: Oxford University Press.

Biber, Douglas & Reppen, Randi (eds). 2015. *The Cambridge Handbook of English Corpus Linguistics*. Cambridge: Cambridge University Press.

Bing, Janet M. & Bergvall, Victoria L. 1996. The question of questions. Beyond binary thinking. In *Rethinking Language and Gender Research. Theory and Practice*, Victoria L. Bergvall, Janet M. Bing & Alice F. Freed (eds), 1–30. London: Longman.

Bogetic, Ksenija. 2013. Normal straight gays. Lexical collocations and ideologies of masculinity in personal ads of Serbian gay teenagers. *Gender and Language* 7(3): 333–367.

Butler, Judith. 1990. *Gender Trouble. Feminism and the Subversion of Identity*. New York: Routledge.

Cameron, Deborah. 1998. Dreaming the dictionary: Keywords and corpus linguistics. *Key Words: A Journal of Cultural Materialism* 1: 35–46.

Cameron, Deborah & Kulick, Don. 2003. *Language and Sexuality*. Cambridge: Cambridge University Press.

Carroll, David & Kowitz, Johanna. 1994. Using concordancing techniques to study gender stereotyping in ELT textbooks. In *Exploring Gender. Questions and Implications for English Language Education*, Jane Sunderland (ed), 73–82. New York: Prentice Hall.

Culpeper, Jonathan. 2009. Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's Romeo and Juliet. *International Journal of Corpus Linguistics* 14(1): 29–59.

Davies, Mark. 2009. The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics* 14(2): 159–190.

Davies, Mark. 2010. The corpus of contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing* 25(4): 447–464.

Davis, Jenny L., Zimman, Lal & Raclaw, Joshua. 2014. Opposites attract: Retheorizing binaries in language, gender, and sexuality. In *Queer Excursions: Retheorizing Binaries in Language, Gender, and Sexuality*, Lal Zimman, Jenny L. Davis & Joshua Raclaw (eds), 1–12. Oxford: Oxford University Press.

Egbert, Jesse & Baker, Paul. 2016. Research synthesis. In *Triangulating Methodological Approaches in Corpus Linguistic Research*, Paul Baker & Jesse Egbert (eds), 183–208. London: Routledge.

Ehrlich, Susan. 2001. *Representing Rape: Language and Sexual Consent*. London: Routledge.

Findlay, Jamie Y. 2017. Unnatural acts lead to unconsummated marriages: Discourses of homosexuality within the House of Lords debate on same-sex marriage. *Journal of Language and Sexuality* 6(1): 30–60.

Gales, Tammy 2015. The stance of stalking: A corpus-based analysis of grammatical markers of stance in threatening communications. *Corpora* 10(2): 171–200.

Grimm, Anne. 2008. *'Männersprache' - 'Frauensprache'? Eine korpusgestützte empirische Analyse des Sprachgebrauchs britischer und amerikanischer Frauen und Männer hinsichtlich Geschlechtsspezifika*. Hamburg: Dr. Kovac.

Hall, Kira. 2013. 'It's a hijra!'. Queer linguistics revisited. *Discourse & Society* 24(5): 634–642.

Harrington, Kate. 2008. Perpetuating difference? Corpus linguistics and the gendering of reported dialogue. In *Gender and Language Research Methodologies*, Kate Harrington, Lia Litosseliti, Helen Sauntson & Jane Sunderland (eds), 85–102. Basingstoke: Palgrave Macmillan.

Jiménez Catalán, Rosa Maria & Ojeda Alba, Julieta. 2008. The English vocabulary of girls and boys. Evidence from a quantitative study. In *Gender and Language Research Methodologies*, Kate Harrington, Lia Litosseliti, Helen Sauntson & Jane Sunderland (eds), 103–115. Basingstoke: Palgrave Macmillan.

King, Brian W. 2009. Building and analysing corpora of computer-mediated communication. In *Contemporary Corpus Linguistics*, Paul Baker (ed), 301–320. London: Continuum.

King, Brian W. 2015. Investigating digital sex talk practices: A reflection on corpus-assisted discourse analysis. In *Discourse and Digital Practices: Doing Discourse Analysis in the Digital Age*, Rodney H. Jones, Alice Chik & Christoph A. Hafner (eds), 130–143. London: Routledge.

Kjellmer, Göran. 1986. 'The lesser man': Observations on the role of women in modern English writings. In *Corpus Linguistics II. New Studies in the Analysis and Exploitation of Computer Corpora*, Jan Aarts & Willem Meijs (eds), 163–176. Amsterdam: Rodopi.

Kulick, Don. 2005. The importance of what gets left out. *Discourse Studies* 7(4/5): 615–624.

Leap, William L. 2015. Queer linguistics as critical discourse analysis. In *The Handbook of Discourse Analysis*, Deborah Tannen, Heidi E. Hamilton & Deborah Schiffrin (eds), 661–680. Chichester: Wiley-Blackwell.

Leipold, Ute. 2006. Constructing the self and constructing the (significant) other. A corpus-based critical analysis of gender identities in personal ads. In *Planing [sic], Gluing and Painting Corpora. Inside the Applied Corpus Linguist's Workshop*, Bernhard Kettemann & Georg Marko (eds), 151–174. Frankfurt am Main: Peter Lang.

Love, Robbie & Baker, Paul 2015. The hate that dare not speak its name? *Journal of Language Aggression and Conflict* 3(1): 57–86.

Manning, Elizabeth. 1997. Kissing and cuddling. The reciprocity of romantic and sexual activity. In *Language and Desire. Encoding Sex, Romance and Intimacy*, Keith Harvey & Celia Shalom (eds), 43–59. London: Routledge.

Marchi, Anna & Taylor, Charlotte. 2009. If on a winter's night two researchers… A challenge to assumptions of soundness of interpretation. *Critical Approaches to Discourse Analysis across Disciplines* 3(1): 1–20.

Marko, Georg. 2006. '… was an incredibly sexy Latin stud.' Critically analysing descriptors in a (pornography) corpus. In *Planing [sic], Gluing and Painting Corpora. Inside the Applied Corpus Linguist's Workshop*, Bernhard Kettemann & Georg Marko (eds), 175–203. Frankfurt am Main: Peter Lang.

Marko, Georg. 2008. *Penetrating Language. A Critical Discourse Analysis of Pornography*. Tübingen: Gunter Narr.

Mautner, Gerlinde. 2009. Corpora and critical discourse analysis. In *Contemporary Corpus Linguistics*, Paul Baker (ed), 32–46. London: Continuum.

Mautner, Gerlinde. 2016. Checks and balances: How corpus linguistics can contribute to CDA. In *Methods of Critical Discourse Studies*, Ruth Wodak & Michael Meyer (eds), 154–179. Los Angeles: Sage.

McEnery, Tony & Baker, Helen. 2017. *Corpus Linguistics and 17th-Century Prostitution: Computational Linguistics and History*. London: Bloomsbury.

McEnery, Tony & Hardie, Andrew. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

Milani, Tommaso M. 2013. Are 'queers' really 'queer'? Language, identity and same-sex desire in a South African online community. *Discourse & Society* 24(5): 615–633.

Mills, Sara. 2008. *Language and Sexism*. Cambridge: Cambridge University Press.

Morrish, Liz & Sauntson, Helen. 2007. *New Perspectives on Language and Sexual Identity*. Basingstoke: Palgrave Macmillan.

Morrish, Liz & Sauntson, Helen. 2011. Gender, desire and identity in a corpus of lesbian erotica. In *Queering Paradigms II. Interrogating Agendas*, Burkhard Scherer & Matthew Ball (eds), 63–81. Frankfurt am Main: Peter Lang.

Motschenbacher, Heiko. 2010. *Language, Gender and Sexual Identity. Poststructuralist Perspectives*. Amsterdam: John Benjamins.

Motschenbacher, Heiko. 2011. Taking Queer Linguistics further. Sociolinguistics and critical heteronormativity research. *International Journal of the Sociology of Language* 212: 149–179.

Motschenbacher, Heiko. 2013. Gentlemen before ladies? A corpus-based study of conjunct order in personal binomials. *Journal of English Linguistics* 41(3): 212–242.

Motschenbacher, Heiko. 2014. Focusing on normativity in language and sexuality studies. Insights from conversations on objectophilia. *Critical Discourse Studies* 11(1): 49–70.

Motschenbacher, Heiko. 2016. A corpus linguistic study of the situatedness of English pop song lyrics. *Corpora* 11(1): 1–28.

Motschenbacher, Heiko. 2018. Sexuality in critical discourse studies. In *The Routledge Handbook of Critical Discourse Studies*, John Flowerdew & John E. Richardson (eds), 388–402. London: Routledge.

Motschenbacher, Heiko. forthcoming a. Language and sexual normativity. In *The Oxford Handbook of Language and Sexuality*, Rusty Barrett & Kira Hall (eds). Oxford: Oxford University Press.

Motschenbacher, Heiko. forthcoming b. *Gay people* and *homosexual persons*: A co-occurrence analysis of sexual descriptive adjectives in COCA.

Murphy, Bróna. 2010. *Corpus and Sociolinguistics. Investigating Age and Gender in Female Talk*. Amsterdam: John Benjamins.

O'Keeffe, Anne & Breen, Michael J. 2007. At the hands of the brothers. A corpus-based lexico-grammatical analysis of stance in newspaper reporting of child sexual abuse cases. In *The Language of Sexual Crime*, Janet Cotterill (ed), 217–236. Basingstoke: Palgrave Macmillan.

O'Keeffe, Anne & McCarthy, Michael (eds). 2012. *The Routledge Handbook of Corpus Linguistics*. London: Routledge.

Partington, Alan. 2014. Mind the gaps: The role of corpus linguistics in researching absences. *International Journal of Corpus Linguistics* 19(1): 118–146.

Pearce, Michael. 2008. Investigating the collocational behaviour of MAN and WOMAN in the BNC using Sketch Engine. *Corpora* 3(1): 1–29.

Rayson, Paul. 2008. From key words to key semantic domains. *International Journal of Corpus Linguistics* 13(4): 519–549.

Rayson, Paul, Leech, Geoffrey & Hodges, Mary. 1997. Social differentiation in the use of English vocabulary. Some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics* 2(1): 133–152.

Sauntson, Helen & Sundaram, Vanita. 2016. Discursive silences: Critically analysing the presence/absence of sexual diversity in the sex and relationships education guidance for England and Wales. In *Global Perspectives and Key Debates in Sex and Relationships Education: Addressing Issues of Gender, Sexuality, Plurality and Power*, Vanita Sundaram & Helen Sauntson (eds), 100–114. Basingstoke: Palgrave Macmillan.

Schmid, Hans-Jörg. 2003. Do women and men really live in different cultures? Evidence from the BNC. In *Corpus Linguistics by the Lune. A Festschrift for Geoffrey Leech*, Andrew Wilson, Paul Rayson & Tony McEnery (eds), 185–221. Frankfurt am Main: Peter Lang.

Schmid, Hans-Jörg. 2015. Does gender-related variation still have an effect, even when topic and (almost) everything else is controlled? In *Change of Paradigms – New Paradoxes: Recontextualizing Language and Linguistics*, Jocelyne Daems, Eline Zenner, Kris Heylen, Dirk Speelman & Hubert Cuyckens (eds), 327–346. Berlin: De Gruyter Mouton.

Schröter, Melani & Taylor, Charlotte. 2018. Introduction. In *Exploring Silence and Absence in Discourse: Empirical Approaches*, Melani Schröter & Charlotte Taylor (eds), 1–21. Basingstoke: Palgrave Macmillan.

Scott, Mike. 2009. In search of a bad reference corpus. In *What's in a Word-List? Investigating Word Frequency and Keyword Extraction*, Dawn Archer (ed), 79–92. London: Routledge.

Sigley, Robert & Holmes, Janet. 2002. Looking at girls in corpora of English. *Journal of English Linguistics* 30(2): 138–157.

Taylor, Charlotte. 2013. Searching for similarity using corpus-assisted discourse studies. *Corpora* 8(1): 81–113.

Thorne, Lisa. 2013. 'But I'm attracted to women': Sexuality and sexual identity performance in interactional discourse among bisexual students. *Journal of Language and Sexuality* 2(1): 70–100.

Vigo, Francesca. 2015. And what about same sex marriages? A corpus-based analysis of lexical choices and social attitudes. In *Languaging Diversity: Identities, Genres, Discourses*, Giuseppe Balirano & Maria Cristina Nisco (eds), 197–210. Cambridge: Cambridge Scholars Publishing.

Wilson, Andrew. 2012. Using corpora in depth psychology. A trigram-based analysis of a corpus of fetish fantasies. *Corpora* 7(1): 69–90.

Zimman, Lal, Davis, Jenny L. & Raclaw, Joshua (eds). 2014. *Queer Excursions: Retheorizing Binaries in Language, Gender, and Sexuality*. Oxford: Oxford University Press.

Author's address

Heiko Motschenbacher

English Department

Western Norway University of Applied Sciences

Postboks 7030, 5020 Bergen

Norway

heim@hvl.no