Carolin Müller-Spitzer/Jan Oliver Rüdiger

THE INFLUENCE OF THE CORPUS ON THE REPRESENTATION OF GENDER STEREOTYPES IN THE DICTIONARY. A CASE STUDY OF CORPUS-BASED DICTIONARIES OF GERMAN

Abstract Dictionaries are often a reflection of their time; their respective (socio-)historical context influences how the meaning of certain lexical units is described. This also applies to descriptions of personal terms such as *man* or *woman*. Lexicographers have a special responsibility to comprehensively investigate current language use before describing it in the dictionary. Accordingly, contemporary academic dictionaries are usually corpus-based. However, it is important to acknowledge that language is always embedded in cultural contexts. Our case study investigates differences in the linguistic contexts of the use of man and woman, drawing from a range of language collections (in our case fiction books, popular magazines and newspapers). We explain how potential differences in corpus construction would therefore influence the "reality"[1] depicted in the dictionary. In doing so, we address the far-reaching consequences that the choice of corpus-linguistic basis for an empirical dictionary has on semantic descriptions in dictionary entries. Furthermore, we situate the case study within the context of gender-linguistic issues and discuss how lexicographic teams can engage with how dictionaries might perpetuate traditional role concepts when describing language use.

Keywords Gender linguistics; corpus-based lexicography; collocations; lexicography equality; gender equality

1. Have you ever googled 'woman'?

In 2019, the British PR manager Maria Beatrice Giovanardi wrote a blog post titled "Have you ever googled 'woman'?" in which she primarily complained about the description of women in various dictionaries, including lexicographic works by Oxford University Press, e. g. that *filly*, *biddy* or *bitch* are listed as synonyms for *woman*:

> The first search involved googling 'woman synonyms' and boom – an explosion of rampant sexism. I thought to myself, 'What would my young niece think of herself if she read this?' […] Should data about how language is used control how women are defined? Or should we take a step back and, as humans, promote gender equality through the definitions of women that we choose to accept? […] We talked about how the dictionary is the most basic foundation of language and how it influences conversations. Isn't it dangerous for women to maintain these definitions – wof women as irritants, sex objects and subordinates to men? (Giovanardi 2020)

She then started a petition at change.org, which was signed by 30,000 people. Oxford University Press responded by sending Katherine Connor Martin the following statement via The Guardian newspaper: The dictionary editors "are taking the points raised in the peti-

[1] What can be seen as "linguistic reality" is a very complex matter that goes beyond the scope of this paper. When we use in the following the term "linguistic reality", we are aware that texts or corpora are not a "description" or "representation" of this assumed reality, but serve to construct and interpret one possible part of this reality from language use (e. g. simply in reading or in specific work such as lexicography).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

tion very seriously […] As ever, our dictionaries strive to reflect, rather than dictate, language so any changes will be made on that basis. (Flood 2019). Here, reference is made to the descriptive tradition of modern lexicography. But in our view, two questions arise from this statement: a) What is regarded as a basis for the 'reflection of language'? In the tradition of modern corpus-based lexicography, it is the underlying corpus 'base'. But does everything from this corpus base always have to be included in the dictionary? Or should it rather be a curated selection? b) Should language use find its way into the dictionary, even if it could perpetuate gender stereotypes that, at least in part, no longer fit with contemporary ideas of society? Is it acceptable to reproduce racist and sexist attitudes exactly as they are (still) used?

2. "The man's a genius" and "she's really a nice woman": gender stereotypes in dictionaries

Dictionaries are often a reflection of their time, i. e. how they describe the meaning of certain lexical units must always be seen in their respective historical context. They are one of the sources to reflect gender roles (Nübling 2010, p. 594) for the first time, the lexicographic construction of gender in more recent editions of German dictionaries (from 1980 onwards. Consider the following example phrases taken from the entries on *man*, *woman*, *girl* and *boy* in the Cambridge Dictionary, reproducing stereotypical gender concepts:[2]

- "He plays baseball, drinks a lot of beer and generally acts like one of the boys."
- "Steve can solve anything – the man's a genius."
- "She's a really nice woman."
- "Who was that beautiful girl I saw you with last night?"
- "Both girls compete for their father's attention."

We understand stereotypes as thinking in group categories, although we acknowledge that this topic is treated in a much more differentiated way in social psychology:

> Indeed, individuals and groups can be said to be the central facts of society. Without individuals there could be no society, but unless individuals also perceive themselves to belong to groups, that is, to share characteristics, circumstances, values and beliefs with other people, then society would be without structure or order. These perceptions of groups are called stereotypes. (McGarty/Yzerbyt/Spears 2002, p. 1)

Such group descriptions concerning gender can be found in many dictionaries. Very pointedly and amusingly, Luise Pusch has shown this for example phrases of the German Duden dictionary of meanings from 1970:[3] The man, i. e. "he", "shows an acrobatic mastery of his body", "his soul is able to encompass the universe" and "great effect emanated from him". "She," on the other hand, "is always neatly dressed," "took the baby out daily," "awaits his return with great anxiety," and "she looked up to him as to a god."[4] Pusch summarizes: "In the preface, the editors write that the 'basic vocabulary of German in its basic meanings' is

[2] https://dictionary.cambridge.org/de/.

[3] Duden Bedeutungswörterbuch, Mannheim 1970.

[4] Original: Der Mann, also „er", „zeigt eine akrobatische Beherrschung seines Körpers", „seine Seele vermag das All zu umfassen" und „große Wirkung ging von ihm aus". „Sie" dagegen „ist immer adrett gekleidet", „hat das Baby täglich ausgefahren", „erwartet mit großer Angst seine Rückkehr" und „sie sah zu ihm auf wie zu einem Gott".

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

to be presented. They succeed in much more: they convey a deep, unforgettable insight into the soul of German, into its basic treasure of feelings and thoughts." (Pusch 1984, p. 144 [own translation]; cf. in more detail on various dictionaries of German Nübling 2010). This may illustrate that dictionaries are often a mirror of their time and thus also one of the important "platforms for productions of gender" (Nübling 2010, p. 594). Similarly, in their analysis of a contemporary Chinese dictionary, Hu/Xu/Hao (2019) point out that

> Women are often constructed in peripheral and domestic roles, as daughter, mother or grandmother. Their experiences are mostly restricted to themselves and their adjacent environment. When they act, their actions rarely bring noticeable changes to other participants or to the environment. Women are described as sensitive, loving and emotional, particularly preoccupied with familial, marital and domestic matters. On the other hand, men are mostly constructed in their central and social roles, as the prototypical adult men. […] Men are described as strong in physical strength, versatile in skills and noble in their actions. In other words, men are represented as valuable, active social members. (Hu/Xu/Hao 2019, p. 28)

Regardless of whether one sees this as an adequate description of 'reality' or as an overly stereotypical representation of men and women, the question arises whether such representations of gender in dictionaries are or can be intentional. For example, John Sinclair states in the preface to the 1987 Collins Cobuild English Language Dictionary that they "have abandoned the convention whereby *he* was held to refer to both men and women." This was done for various reasons, including the fact that "it is a very sensitive matter for those who have pointed out the built-in sexism of English" (Sinclair 1992, p. XX). This conscious positioning is particularly relevant for dictionaries because they can be understood as normative instances, even if they are primarily intended to be descriptive:

> This brings up the question of usage and authority. These concepts must support each other or no-one will respect either of them. If their close relationship breaks down, and authority is not backed up by usage, then no-one will respect it. […] Similarly, no-one will respect usage if it is merely an unedited record of what people say and write. […] Any successful record of a language such as a dictionary is itself a contribution to authority. (Sinclair 1992, compare also: Ripfel 1989, p. 204; Barnickel 1999, p. 171; Hidalgo Tenorio 2000, p. 225; Kotthoff/Nübling 2018, p. 180)

Against this background, lexicographers have a special responsibility. After Pusch's essay cited above, attempts were made in the Duden editorial office to improve the dictionary in many areas, e. g. to avoid unnecessarily stereotypical example phrases and to systematically include female occupational designations when they are common. (Kunkel-Razum 2004; for general comments, see Westveer/Sleeman/Aboh 2018). The main point here is to express awareness of the issue:

> Of course, dictionaries are not supposed to "straighten out" asymmetrical conditions that are solidified in the language system. It is undisputed and anchored in the German language (in the lexicon) that the entry *girl* always has to refer to the *easy girl* and the entry *boy* to the *tough boy*. It is not a matter of demanding a *heavy girl* or a *light boy* […]. Neither is it about *pregnant men* and *female machos*. It is about lexicographic doing gender. […] the question of which position on a scale from undoing gender via doing gender to hyper-ritualized gender the dictionaries take, in other words, which "degree of dramatization" they adopt – and whether they possibly engage in such dramatization themselves. (Nübling 2010, p. 595 [own translation])

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

The representation of gender in dictionaries thus seems to be caught between language use and lexicographic-moral responsibility. In our paper, in addition to discussing how much the lexicographer must or should intervene in the description of language use, we first investigate whether language use is indeed uniform at all. This question is particularly pertinent because we discovered strongly stereotypical statements about men and women in the entries of a modern corpus-based dictionary of German. It was newly compiled and therefore did not contain any old example phrases, e. g. examples inherited from earlier editions or other, older dictionaries. This finding was our starting point to examine the question of the data basis of 'language use' reflected in dictionaries.

3. Case study: influence of the corpus base on collocation sets for *Mann* (*man*) and *Frau* (*woman*)

3.1 Current lexicographic practice in German dictionaries

The starting point of our case study is the observation that even in modern corpus-based dictionaries of German, e. g. *elexiko*,[5] the descriptions of entries such as *Mann* or *Frau* are more influenced by stereotypes than we expected. *elexiko* is compiled from contemporary sources and does not contain old examples. This is why we thought we might find a more 'modern' representation of *Mann* or *Frau* in the dictionary. However, this is not the case.

In *elexiko*, collocation sets are listed for each head word. In the case of *Mann* and *Frau*, selecting the most frequent collocators leads to strongly different representations. It is particularly striking that for *Mann*, the agent role constitutes the second collocation set ("What does a *man* do?"), whereas for *Frau*, the patient role ("What happens to a *woman*?") is listed second – an imbalance that some researchers have already criticized as 'doing' gender (Hidalgo Tenorio 2000; Nübling 2010; Hu/Xu/Hao 2019) as how bias itself may organize human beings' experience by means of language in use. There exist well-known cultural stereotypes associated with the male and female conditions, and it is necessary to acknowledge the limitations to the application of many an impressionistic linguistic study on such issues. Taking this into account, the aim of this paper is to look at the way certain aspects of present-day English (a natural-gendered language). The fact that these collocation sets are presented in the dictionary in this way is due to the frequency of the groups, i. e. the patient role of women is much more prominent in the corpus base of *elexiko* than the agent role. For men, it is the other way round. Within the collocation sets for "what is discussed in connection with *man* or *woman*?", *man* collocates with: *car*, *erectile dysfunction*, *fire department*, *soccer*, *equality*, and *handball*. For *woman*, it is *age*, *occupation*, *breast cancer*, *emancipation*, *employment*, *birth*, *children*, *sex*, and *menopause*.

The *elexiko* team expresses critical awareness of these stereotypical representations. They point out that in the case of *woman*, reference is often made to their social roles in the family context (*single parent*, *divorced*, *unmarried*) or their general employment status (*unemployed*, *employed*), whereas in the case of *man*, such characterisations are absent. Adjectives such as *armed*, *masked*, *suspicious*, *hooded* only appear in the entry *man*, probably because the newspaper-heavy corpus contains many reports of violence and crime (Klosa/Storjohann 2011, p. 64).

[5] elexiko (2003 ff.), in: OWID – Online Wortschatz-Informationssystem Deutsch. Ed. by Leibniz-Institut für Deutsche Sprache, Mannheim, http://www.owid.de/wb/elexiko/start.html.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Further stereotypical representations can be found in the computer-generated collocation profiles ("Typische Verbindungen (computergeneriert)") within the Duden Online website.[6] Typical adjectives for *man* are *young*, *old*, *rich*, *strong*, *adult*, *powerful*, *armed* and *right*, whereas the typical ones for *woman* are *young*, *old*, *beautiful*, *tall*, *naked*, *pregnant*, *gracious* and *employed*. The corpora on which the two dictionaries (*elexiko*, Duden Online) are based – like the large linguistic corpora on German in general – are dominated by newspaper texts and such corpora have already been criticised as unbalanced in the context of lexicography (cf. Rundell/Atkins 2013, p. 1339). This is particularly relevant in the case of the computer-generated collocation profiles in Duden because they are obtained from the current newspaper-heavy Duden corpus. The extent to which this corpus base influences the representations must be carefully scrutinized. Contrastingly, the example phrases are more strongly informed by earlier editions of the dictionary and manual lexicographic analysis.

Linguistic practice is always embedded in a cultural context. "Language exists only in its use, and this is always culturally framed; at the same time, cultural facts, cultural habits, conceptualizations, and values are constructed and sedimented – indeed, archived – through language and in language" (Günthner/Linke 2006, p. 19, own translation). The empirical basis of lexicographic work transfers this linguistic-cultural context, and therefore a particular perspective on the world, into the dictionary. For example, men may indeed be more criminal than women, and women are also raped by men – however, it is open to debate whether exactly these aspects of 'reality' should be the main perspective of dictionaries.

The German collocation dictionary provides another solution for this issue: The entries for *Mann* and *Frau*, as with *elexiko* and Duden, are also designed to represent language use, but they are clearly displayed in a parallel structure (see Fig. 1). This approach requires more manual post-editing of the corpus data (which might have other, also negative, implications). According to a colleague who worked on the dictionary, this was a conscious decision.

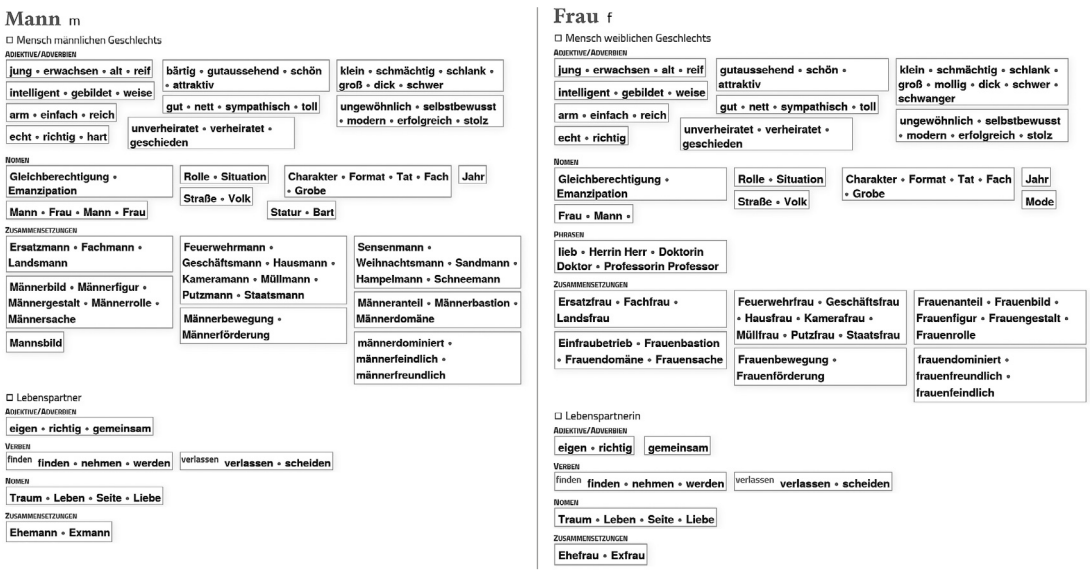

Fig. 1: Entries *Mann* (*man*) and *Frau* (*woman*) in the German collocation dictionary (Buhofer et al. (2015)

[6] www.duden.de.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

In a next step, we present a case study in which we investigate whether the collocation sets for *Mann* and *Frau* would change significantly if the corpus base was not predominantly composed of newspaper texts. We examine whether different corpus bases lead to different embeddings of *Mann* and *Frau*, addressing the urgent question of what we consider to "reflect language". We then discuss which methodological implications this could have for corpus-based lexicography in general. We end by addressing the fundamental question of how lexicographers should or could position themselves regarding the representation and perpetuation of gender stereotypes in dictionaries.

3.2 Method

The analyses presented in the following are based on three corpora constructed from different source materials:

- The corpus 'Fiction Books'[7] is based, as the name suggests, on various works of fiction (20th and 21st century). These corpora are listed in DeReKo[8] with the prefix 'LOZ-*'. Additionally, the corpora 'Mannheimer Korpus 1' and the 'THM – Thomas Mann Korpus' are included because they also consist of fictional texts.
- The '*elexiko*' corpus is based on the sources used for the '*elexiko*' dictionary (only newspaper texts), as well as more-recent newspaper-based documents (up to DeReKo Release 2021-I). Sources are: *St. Galler Tagblatt*, *Berliner Zeitung*, *Braunschweiger Zeitung*, *Burgenländische Volkszeitung*, *Bonner Zeitungskorpus*, *Deutsche Presse-Agentur*, *Tages-Anzeiger*, *Frankfurter Allgemeine*, *Handbuchkorpus*, *Hannoversche Allgemeine*, *Hamburger Morgenpost*, *Tiroler Tageszeitung*, *Kleine Zeitung*, *Berliner Morgenpost*, *Mannheimer Morgen*, *Salzburger Nachrichten*, *Niederöstereichische Nachrichten*, *Die Presse*, *Frankfurter Rundschau*, *Rhein-Zeitung*, *Der Spiegel*, *Die Südostschweiz*, *die tageszeitung*, *Vorarlberger Nachrichten*, *Oberösterreichische Nachrichten* and *Die Zeit*.
- The 'magazines' corpus consists of various periodical magazines. Sources are: *art*, *BEEF!*, *brand eins*, *BRIGITTE*, *Capital*, *Chefkoch*, *Couch*, *Eltern*, *Essen und Trinken*, *Gala*, *GEO*, *Living at Home*, *Nido*, *NEON*, *Psychologie Heute* and *Schöner Wohnen*.

	fiction books	magazines	newspapers (*elexiko*)
Time Range	1893–2011	2005–2020	1947–2020
Texts	1.320	60.066	15.831.499
Sentences	1.221.373	2.511.280	263.625.222
Tokens	22.132.897	37.771.792	4.398.207.319

Table 1: The three differently confected corpora for our case study

[7] One reviewer of the abstract correctly pointed out that we compare text types (newspapers, magazines) with a genre (fiction). Of course, fictional texts can also be found in newspaper texts, even if only to a small extent. However, we think that it is legitimate for our case study to proceed in this way, because corpus compilations in corpus linguistic practice usually include whole sources: whole newspapers, whole magazines or whole books. However, calling the fictional corpus a "book corpus" because we took fiction books seemed too general. By calling it "fiction books", however, we hope to have appropriately taken up the criticism.

[8] Cf. Kupietz et al. (2010, 2018).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

As can be seen in Table 1, the three corpora differ considerably, both in terms of the number of texts and the number of tokens. The three corpora also encompass different time periods. For the fiction books corpus, older texts were included a) because there are very few fiction books in the IDS corpora in general, and b) because limiting the corpus to recent texts would have resulted in too small a collection for the analyses required. The popular magazines such as Beef!, Brigitte Woman, Chefkoch or Living at Home are very recent, dating only from 2005 up to 2020. The *elexiko* corpus spans are wider time frame, namely from 1947 to 2020, but the largest amount of *elexiko* corpus texts can also be assigned to a very similar time period as the magazines. In the following, the *elexiko* corpus is referred to only as the newspaper corpus, since it consists exclusively of newspapers.

The corpora were imported into CorpusExplorer (Rüdiger 2021). For each search term (*Mann*/*Frau*), the corpora were separated so that only texts containing the particular search term were used for the co-occurrence calculation. A token-span (limit) for the calculation was not specified, and there was no restriction on parts of speech (POS). The sentence boundary was used to identify co-occurrences. Common co-occurrences were filtered by the 100 most significant entries (based on Poisson distribution, (cf. Heyer/Quasthoff/Wittig 2006, p. 134). To avoid visual distortion, we have filtered out the co-occurrences *young* and *old*, as their inclusion makes the observation and interpretation of the tag clouds more difficult.

3.3 Results

Figure 2 shows the most significant co-occurrences to *Mann* and *Frau* as a result of our analyses.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Fig. 2: All co-occurrences for *Mann* and *Frau* (the font size correlates with the significance based on Poisson distribution)

For the following comparison, the co-occurrences were filtered to include only tokens that were annotated at least once by the TreeTagger (cf. Schmid 1995) with the specific POS tag (here adjectives). In Figure 3 we see all adjectives co-occurring with *Mann* and *Frau* in our three corpora.

Fig. 3: Adjective co-occurrences for *Mann* and *Frau* (the font size depends on the significance based on Poisson-distribution)

Adjectives are used, among other things, to describe people. Thus, they would be included in a collocation set like "What is a *woman* or a *man* like?". What tendencies do our three corpora show in this regard?

Women are described in the fiction books partly regarding their external characteristics: *blond*, *pretty* or *attractive* (*blond*, *hübsch*, *attraktiv*), but also in terms of their marital status

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

(*married* or *divorced* – *verheiratet*, *geschieden*), or even *pregnant* (*schwanger*). The adjective *schweigestill* seems to us to be a tagging error (it is not a German adjective) and *gnädig* rather points to the quasi-lexicalized address *gracious woman* (*gnädige Frau*). In the newspaper texts, women are also described as *pregnant* (*schwanger*) or *working* (*berufstätig*), but also as being *raped* (*vergewaltigt*). Even in this case, however, passive constructions with participle 2 uses may be mixed with adjective uses. *Sexual* (*sexuell*) and *affected* (*betroffen*) could point to usage contexts such as *affected by sexual violence*. In the magazine texts, women are *self-confident*, *employed*, *attractive*, *pregnant*, *independent*, or *emancipated* (*selbstbewusst*, *berufstätig*, *attraktiv*, *schwanger*, *unabhängig*, *emanzipiert*). The magazines corpus is therefore the only dataset in which women are characterized by significant collocators that have nothing to do with their appearance or social role. We cannot efficiently classify the adjectives *feminine* (*weiblich*) and *masculine* (*männlich*); they are probably not used as direct attributes for women. The significant co-occurrence *sexual* is again likely to occur in bigger phrasal chunks, just like in the newspaper texts. What can be seen overall is that the collocation sets, as they would then be listed in the dictionary, would differ visibly depending on the corpus base.

The differences between corpora become even clearer with the adjectives co-occurring with *man*: in the fiction books, descriptive adjectives such as *gaunt*, *stout*, *stocky*, *bearded* or *lanky* (*hager*, *kräftig*, *untersetzt*, *bärtig*, *schmächtig*) dominate. *Dressed* (*gekleidet*) may not always be used as a direct attribute. In the newspaper texts, violent acts are a predominant topic. Logically, they are discussed more frequently in newspapers due to their news value: *armed*, *masked*, *alcoholized*, *previously convicted* (*bewaffnet*, *maskiert*, *alkoholisiert*, *vorbestraft*), but also more general words like *unemployed* or *powerful* (*arbeitslos*, *mächtig*). In magazines, men are described as *attractive*, *married*, *bearded*, *naked*, *gay* or **-looking* (*attraktiv*, *verheiratet*, *bärtig*, *nackt*, *schwul*, *aussehend*). Surprisingly, a considerable number of terms related to appearance, social role or sexual orientation are found here. The examples show clearly how differently 'linguistic reality' turns out, depending on which empirical basis is used.

Fig. 4: Verbal co-occurrences for *Mann* (the font size depends on the significance based on Poisson-distribution)

Similar differences appear in the verbal co-occurrences for *Mann*, i. e. fillers to collocation sets like "What does a *man* do?" or "What happens to a *man*?". Verbs in the fictional books are *muster*, *marry*, *observe*, *sit opposite*, *turn to* (*mustern*, *heiraten*, *beobachten*, *gegenübersitzen*, *zuwenden*; *erwachsen* is again a tagging error, *untersetzen* presumably also). In the newspaper texts, the context of violence is again predominant: *arrest*, *assault*, *threaten*, *shoot*, and *rape* (*festnehmen*, *überfallen*, *bedrohen*, *erschießen*, *vergewaltigen*) are particularly significant co-occurrences. In magazines, again, words referring to love life, money or power status are frequent collocators: *marry*, *fall in love*, *question*, *earn*, *cheat*, *question*, or *dominate* (*heiraten*, *verlieben*, *befragen*, *verdienen*, *betrügen*, *befragen*, *dominieren*). Again, the linguistic reality differs greatly.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Fig. 5: Nominal co-occurrences for *Frau* (the font size depends on the significance based on Poisson-distribution)

As a final example, we examine the nominal co-occurrences for *Frau*, i. e., "What is the topic of discussion in connection with *woman*?". For newspapers, the answer would be: *child*, *husband*, *violence*, *equality*, *social service* (*Kind*, *Ehemann*, *Gewalt*, *Gleichberechtigung*, *Sozialdienst*). For magazines, on the other hand: *leadership position*, *financial advisor*, *study*, *equality* (*Führungsposition*, *Finanzberaterin*, *Studie*, *Gleichberechtigung*; *percent* is more likely to be part of a phrase like "x percent of women are …."). Reflecting on language use would thus lead to very different results depending on the linguistic-thematic embedding of the words in the various text groups.

One should always keep in mind that co-occurrences say little about frequencies, but more about the strength of a connection. The fact that *woman* is so strongly associated with *gracious* (*gnädig*) in fiction does not mean that gracious women are often mentioned in total numbers, but that a (presumably low-frequency) word like *gracious* has a significant affinity to *woman*. Co-occurrences therefore indicate that certain activities or characteristics are strongly associated with women or men in the texts, which is more interesting for corpus-based research than mere frequencies.

3.4 Discussion and methodological implications

Our results show that in the newspaper texts, the common features of *women* and *men* as people who share many characteristics and actions step back in favour of the differences. The context of violence, for example, which is particularly over-represented in the *elexiko* entries,[9] is dominant only in the newspaper corpus. This is one of the instances where it becomes clear that the corpus basis can bring an unnecessarily strong bias towards *doing gender* into the dictionary (cf. also Nübling 2010, p. 620). This is especially problematic for lexicography:

> In fact, the question is to what extent a dictionary can involve a linguistic change; or, simply, whether its role in that process must be only one of perpetuation of what is actually supported by textual evidence; in other words, why a dictionary is allowed to repeat values which imply a biased representation of reality […]. (Hidalgo Tenorio 2000, p. 227)

Even if one assumes that a linguistic perspective always contains a "biased representation of reality", the case study has shown that the lexicographer chooses one of these linguistic views by selecting a specific corpus base, and that these linguistic perspectives on 'reality'

[9] In the entry *man* in *elexiko*, the first three verbal co-occurrences are *dominate*, *murder* and *shoot*.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

differ greatly. Gender stereotypes appear to be particularly strong in newspaper texts. These differences do not exist 'per se':

> There are not "the" gender differences in reality. […] This is neither to straighten out nor to idealize real relations nor to practice political correctness, but simply not to take a position on certain points – just as dictionaries do not take a position on racisms and anti-Semitisms (which can be found in reality as well as in corpora) by not reproducing them. (Nübling 2010, p. 628 [own translation])

In our opinion, it needs to be investigated more closely and discussed more intensively which implications go along with these findings. Our results show that different corpus texts lead to different linguistic representations of men and women, and that it should be best-practice to build dictionary entries on a diversified empirical base. However, more stratified compilation of the corpus may not be the best solution either, because it is then no longer possible to distinguish the different influences of the individual text groups. One possibility might be to at least refine the methods for analyzing vocabulary for a general dictionary, e.g. by performing co-occurrence analyses with different corpora containing different text types, and then comparing the resulting lists. This approach, according to our case study, is more likely to result in the most diverse representation possible. It would then also be possible to draw more precise conclusions about which texts have which influences. Our approach follows Sinclair's clarion callfor a very fine-grained documentation of all corpus data in order to be able to better interpret the results of corpus analyses:

> Also at any time a researcher may get strange results, counter-intuitive and conflicting with established descriptions. Neither of these factors proves that there is something wrong with the corpus, because corpora are full of surprises, but they do cast doubt on the interpretation of the findings, and one of the researcher's first moves on encountering unexpected results will be to check that there is not something in the corpus architecture or the selection of texts that might account for it. (Sinclair 2004, chap. 1)

Of course, this requires a very good lexicographic working environment, so that such procedures do not become too time-consuming. In any case, it becomes clear that the linguistic-technological methods cannot be used as a 'black box', but must be intellectually understood in order to be able to correctly classify the findings. The lexicographic work environment should make the variability of language use explorable.

4. Concluding remarks

Even though the orientation to actual language use in dictionary writing is certainly a very important principle of modern lexicography that has made dictionaries better tools, we believe that orientation to language use does not relieve lexicographers of their responsibility to take the political or social implications that language descriptions may have into account. As Nübling puts it: "Overall, one should assume that there is an awareness of gender constructions, especially in lexicographic teams, at the turn of the 20th and 21st centuries." (Nübling 2010, p. 609). A good compromise is certainly first to research language use with as much reflection (and self-reflection) as possible and then also – as one does with offensive or vulgar expressions – to find a compromise between language use orientation and the handing-down of outdated role models. We want to end with 'food for thought', citing David Foster Wallace's essay on "Authority and American Usage" in which he formulates the weak points of descriptive lexicography somewhat provocatively:

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

> But these flaws still seem awfully easy to find. Probably the biggest one is that the Descriptivists' "scientific lexicography" – under which, keep in mind, the ideal English dictionary is basically number-crunching: you somehow observe every linguistic act by every native/naturalized speaker of English and put the sum of all these acts between two covers and call it The Dictionary – involves an incredibly crude and outdated understanding of what *scientific* means. It requires a naïve belief in scientific Objectivity, for one thing. Even in the physical sciences, everything from quantum mechanics to Information Theory has shown that an act of observation is itself part of the phenomenon observed and is analytically inseparable from it. (Wallace 2001, p. 46)

References

Barnickel, K.-D. (1999): Political correctness in learners' dictionaries. In: Herbst, T./Popp, K. (eds.): The perfect learners' dictionary (?). Berlin/Boston, pp. 161–174. http://doi.org/10.1515/9783110947021.161.

Buhofer, A. H. et al. (2015): Feste Wortverbindungen des Deutschen. Kollokationenwörterbuch für den Alltag. In: Neuphilologische Mitteilungen 116 (1), pp. 242–244. https://www.jstor.org/stable/26372470.

Flood, A. (2019): Thousands demand Oxford dictionaries 'eliminate sexist definitions'. In: The Guardian, 17 September. https://www.theguardian.com/books/2019/sep/17/thousands-demand-oxford-dictionaries-eliminate-sexist-definitions (last access: 22-03-2022).

Giovanardi, M. B. (2020): Open letter calling on @OxUniPress to change their entry for the word "woman" #SexistDictionary. Change.org. https://www.change.org/p/change-oxford-dictionary-s-sexist-definition-of-woman/u/25841171 (last access: 17-03-2022).

Günthner, S./Linke, A. (2006): Linguistik und Kulturanalyse – Ansichten eines symbiotischen Verhältnisses/Linguistics and cultural analysis – aspects of a symbiotic relationship. In: Zeitschrift für Germanistische Linguistik (ZGL) 34 (1–2), pp. 1–27. http://doi.org/10.1515/ZGL.2006.002.

Heyer, G./Quasthoff, U./Wittig, T. (2006): Text mining: Wissensrohstoff Text: Konzepte, Algorithmen, Ergebnisse. (= IT lernen). Herdecke. http://deposit.dnb.de/cgi-bin/dokserv?id=2783785&prov=M&dok_var=1&dok_ext=htm.

Hidalgo Tenorio, E. (2000): Gender, sex and stereotyping in the Collins COBUILD English language dictionary. In: Australian Journal of Linguistics 20 (2), pp. 211–230. http://doi.org/10.1080/07268600020006076.

Hu, H./Xu, H./Hao, J. (2019): An SFL approach to gender ideology in the sentence examples in the Contemporary Chinese Dictionary. In: Lingua 220, pp. 17–30. http://doi.org/10.1016/j.lingua.2018.12.004.

Klosa, A./Storjohann, P. (2011):): Neue Überlegungen und Erfahrungen zu den lexikalischen Mitspielern. In: Klosa, A. (ed.): *elexiko*. Erfahrungsberichte aus der lexikografischen Praxis eines Internetwörterbuchs. (= Studien zur Deutschen Sprache 55). Tübingen, pp. 49–80. https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/5154.

Kotthoff, H./Nübling, D. (2018): Genderlinguistik: Eine Einführung in Sprache, Gespräch und Geschlecht (= Narr Studienbücher). Tübingen.

Kupietz, M./Belica, C./Keibel, H./Witt, A. (2010): The German Reference Corpus DeReKo: A primordial sample for linguistic research. In: Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010), Valletta, Malta. Paris, pp. 1848–1854.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Kupietz, M./Lüngen, H./Kamocki, P./Witt, A. (2018): The German Reference Corpus DeReKo: New developments – new opportunities. In: Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018). Paris, pp. 4353–4360.

McGarty, C./Yzerbyt, V. Y./Spears, R. (2002): Social, cultural, and cognitive factors in stereotype formation. In: McGarty, C./Yzerbyt, V. Y./Spears, R. (eds.): Stereotypes as explanations: The formation of meaningful beliefs about social groups. New York, pp. 1–15. http://doi.org/10.1017/CBO9780511489877.002.

Nübling, D. (2010): Zur lexikografischen Inszenierung von Geschlecht. Ein Streifzug durch die Einträge von Frau und Mann in neueren Wörterbüchern. In: Zeitschrift für Germanistische Linguistik (ZGL) 37 (3), pp. 593–633. http://doi.org/10.1515/ZGL.2009.037.

Ripfel, M. (1989): Die normative Wirkung deskriptiver Wörterbücher. In: Hausmann, F. J. et al. (eds.): Wörterbücher – Dictionaries – Dictionnaires. Ein Internationales Handbuch zur Lexikographie. (= Handbücher zur Sprach- und Kommunikationswissenschaft 5.1). Berlin/New York, pp. 189–207.

Rüdiger, J. O. (2021): CorpusExplorer. Düsseldorf. http://corpusexplorer.de.

Rundell, M./Atkins, B. T. S. (2013): Criteria for the design of corpora for monolingual lexicography. In: Gouws, R. H. et al. (eds.): Supplementary volume dictionaries. An international encyclopedia of lexicography, pp. 1336–1343. http://doi.org/10.1515/9783110238136.1336.

Schmid, H. (1995): Improvements in Part-of-Speech Tagging with an Application to German. Proceedings of the ACL SIGDAT-Workshop. Dublin, Ireland.

Sinclair, J. (1992): Introduction. In: Collins cobuild English language dictionary. London, pp. XV–XXI.

Sinclair, J. (2004). "Developing linguistic corpora: a guide to good practice". https://users.ox.ac.uk/~martinw/dlc/chapter1.htm.

Wallace, D. F. (2001): Democracy, English, and the wars over usage. In: Harper's Magazine, pp. 39–58.

Contact information

Carolin Müller-Spitzer
Leibniz-Institut für Deutsche Sprache Mannheim
mueller-spitzer@ids-mannheim.de

Jan Oliver Rüdiger
Leibniz-Institut für Deutsche Sprache Mannheim
ruediger@ids-mannheim.de

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.