ELEXIS TRAVEL GRANT REPORT

Report on Elexis Transnational Research Visit Grant at INSTITUUT VOOR DE NEDERLANDSE TAAL, LEIDEN, THE NETHERLANDS

(Leiden, Netherlands, June 13 – June 22, 2022)

Prof. Dr. Carolin Müller-Spitzer

Project title: How is the corpus influencing collocation sets in dictionaries? Enhancing a study on German collocation sets for *Mann* and *Frau* to a contrastive German-Dutch study

The stay financed by the Elexis Travel grant was used to work together with Carole Tiberius and other colleagues at INT on a contrastive German-Dutch study. It is on the influence of the corpus on collocation sets in dictionaries, especially for the entries *man* and *woman*. The initial study on German will be briefly outlined in the following.

Initial Study

The initial study (Müller-Spitzer & Rüdiger 2022, Müller-Spitzer & Lobin 2022) was about the influence of the corpus on collocation sets in dictionaries, exemplified with the entries *Mann* and *Frau.* The starting point of this study was the observation that even in modern corpus-based dictionaries of German, e.g. elexiko[1], the descriptions of entries such as *man* or *woman* are more influenced by stereotypes than we expected. In elexiko, collocation sets are listed for each keyword. In the case of *Mann* (man) and *Frau* (woman), the selection of the most frequent collocates leads to very different representations. It is particularly striking that in the article *Mann*, the agent role constitutes the second collocation set ("What does a man do?"), whereas in the case of *Frau*, the patient role ("What happens to a woman?") is listed second; an imbalance that some researchers have already criticised as *doing gender* (Nübling 2009; Hu, Xu & Hao 2019; Hidalgo Tenorio 2000). The fact that this is presented in the dictionary in this way is due to the frequency of the groups, i.e. in the case of women, the patient role is much more prevalent in the corpus texts of the elexiko corpus than the

[1] https://www.owid.de/docs/elex/start.jsp.

agent role. For men, it is the other way round. Another example of stereotypical representation of gender roles is Duden Online[2], but only a certain item class, namely the computer-generated collocation profiles. Typical adjectives for *Mann* are *young*, *old*, *rich*, *strong*, *adult*, *powerful*, *armed* and *right*, whereas the typical ones for *Frau* are *young*, *old*, *beautiful*, *tall*, *naked*, *pregnant*, *gracious* and *employed*.

The corpora on which the two dictionaries (elexiko, Duden Online) are based are - like the large linguistic corpora on German in general - dominated by newspaper texts (critical to unbalanced corpora in the context of lexicography cf. Rundell & Atkins 2013: 1339). In our case study for German, we show how the linguistic contexts of *man* and *woman* obtained on the basis of newspaper texts differ from other samples, e.g. texts of fiction or popular magazines, and how different the 'reality' shown in the dictionary would look if the corpus was composed differently.

One example are the verbal co-occurrences for *Mann* (cf. Fig. 1), i.e. fillers to collocation sets like "What does a *man* do?" or "What happens to a *man*?". Verbs in the fictional books are *scrutinize, marry, observe, sit opposite, turn to (mustern, heiraten, beobachten, gegenübersitzen, zuwenden*). In magazines, words referring to love life, money or power are frequent collocators: *marry, fall in love, question, earn, cheat,* or *dominate* (*heiraten, verlieben, befragen, verdienen, betrügen, , dominieren)*. In the newspaper texts, the context of violence is predominant: *arrest, assault, threaten, shoot,* and *rape* (*festnehmen, überfallen, bedrohen, erschießen, vergewaltigen*) are particularly significant co-occurrences. Accordingly, the 'linguistic reality' differs greatly depending on which corpus we chose to analyse collocations (and so would do the lexicographic entries).

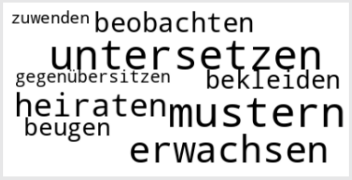
VV	fiction books	magazines	newspapers
Mann			

Fig.1: Verbal co-occurrences for *Mann* (the font size depends on the significance based on Poisson-distribution).

Our results for German show that newspaper texts preferably display differences between men and women instead of making common features, characteristics and actions the subject of discussion. The context of violence, for example, which is particularly over-represented in the elexiko entries,[3] is dominant only in the newspaper corpus. It becomes clear that the

[2] http://www.duden.de.

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

corpus basis can bring an unnecessarily strong bias towards *doing gender* into the dictionary (cf. also Nübling 2010: 620). The aim of the joint study during the guest stay at the INT was to explore whether the same bias (newspaper-heavy corpora = context of violence) can also be observed in Dutch corpus collections.

Joint Work in at the INT in Leiden

The stay in Leiden was used to cooperate (with Carole Tiberius and other colleagues there, cf. e.g. Steurs et al. 2021) to enhance the study described above with data from Dutch. The INT was chosen because it has the corpora, the corpus analysis systems and the knowledge necessary for the analyses.

We needed the first time of the guest stay to find comparable corpus resources. First, we took the corpus the Woordcombinaties[4] project uses. It contains contemporary language material that mainly comes from newspapers (NRC and De Standaard, thus texts from the Netherlands and Flanders, from 2012-2018) and consists of just over 230 million words. We took this corpus as a 'newspaper-corpus'. Second, we chose the ANW literature corpus, which is a subcorpus of the ANW corpus[5]. The Corpus of Literary Texts contains essays, novels, stories and drama, both original and translated work. The selection takes into account a balanced spread in time and a reasonable distribution between North (Netherlands) and South (Belgium). It consists of approx. 20 million tokens.

The analyses of the different collocation sets show strikingly similar results as the study for German. In fiction texts the verbal collocates for *vrouw* and *man* are very similar to each other; in the newspaper texts, they differ greatly, and are especially affected by the discourse around violence (cf. Fig 2 & 3).

It is also striking how similar the collocation sets are regarding individual lexical items (German/Dutch, cf. Fig 4). In both languages, we find highly significant verbal collocates like *to arrest* (*festnehmen/arresteren, oppakken, annhouden*), *to judge* (*veroordelen/verurteilen*), *to shoot* (*doodschieten/erschießen*).

[4] https://woordcombinaties.ivdnt.org/.
[5] https://anw.ivdnt.org/anwcorpus.

vrouw	man
	

Fig.2: Verbal co-occurrences for *vrouw* and *man* in Dutch fiction texts.

vrouw	man
	

Fig.3: Verbal co-occurrences for *vrouw* and *man* in Dutch newspaper texts.

Mann	man
	

Fig.4: Verbal co-occurrences for *Mann* and *man* in German and Dutch newspaper texts.

The Woordcombinaties team at INT using the corpus which is dominated by newspaper texts also analysed the lexemes *man* and *vrouw* during the guest stay. Here, too, it can be seen, especially in the 'object-of-relation', that many of the verbal collocates come from the context of violence (n = 15/39; n = 5/15 as victim, n = 10/15 as offender). For comparison: 9 of 14 verbs in elexiko listed as collocation fillers for "What does a man do?" can be assigned to the context of violence (*dominieren, ermorden, erschießen, schießen, (sich) verletzen, sterben, stürzen, töten, vergewaltigen*). We proofed additionally whether the comparison between texts from the Netherlands and texts from Belgium show any major differences. However, this is not the case (cf. Fig. 5). Thus, the contrastive results strongly suggest that it seems to be indeed a newspaper-bias, not a specific feature of the German-language corpora we used for our initial study.

Fig.5: Verbal co-occurrences for *Mann* and *man* in German and Dutch newspaper texts.

It was very interesting to discuss these results with colleagues at INT, both one-on-one and in connection with a talk I gave there, since the results raise several questions, e.g.: i) regarding corpus selection: How suitable are newspaper-heavy corpora as a basis for lexicographic analysis when the context of violence is so particularly present there? ii) Do lexicographers have to intervene when the corpus brings a strong *doing gender* into the dictionary? A good compromise seems to be first to research language use with as much reflection (and self-reflection) as possible, and then - as a lexicographer does with offensive or vulgar expressions - to find a balance between language use orientation and the handing-down of outdated role models.

Perspectives

The analyses seem promising enough for us to continue our joint work and probably submit a paper together for the next eLex conference. However, there are still some details to be clarified, especially regarding the standardisation of the collocation measures and the

comparability of the corpora. Without the cooperation with the INT as a starting point, such a joint study would not have been possible.

Bibliography

Hidalgo Tenorio, Encarnación. 2000. Gender, Sex and Stereotyping in the Collins COBUILD English Language Dictionary. *Australian Journal of Linguistics*. Routledge 20(2). 211–230. https://doi.org/10.1080/07268600020006076.

Hu, Huilian, Hai Xu & Junjie Hao. 2019. An SFL approach to gender ideology in the sentence examples in the Contemporary Chinese Dictionary. *Lingua* 220. 17–30. https://doi.org/10.1016/j.lingua.2018.12.004.

Müller-Spitzer, Carolin & Rüdiger, Jan-Oliver 2022. The influence of the corpus on the representation of gender stereotypes in the dictionary. A case study of corpus-based dictionaries of German. Euralex Proceedings 2022 (forthcoming).

Müller-Spitzer, Carolin & Lobin, Henning. 2022. Leben, lieben, leiden: Geschlechterstereotype in Wörterbüchern, Einfluss der Korpusgrundlage und Abbild der sprachlichen ‚Wirklichkeit'. In Gabriele Diewald/Damaris Nübling (eds.) *Genus, Sexus, Gender - Neue Forschungen und empirische Studien zu Geschlecht im Deutschen* (Reihe Linguistik: Impulse und Tendenzen), S. 35–64.

Nübling, Damaris. 2009. Zur lexikografischen Inszenierung von Geschlecht. Ein Streifzug durch die Einträge von Frau und Mann in neueren Wörterbüchern. De Gruyter 37(3). 593–633. https://doi.org/10.1515/ZGL.2009.037.

Rundell, Michael & Beryl T. Sue Atkins. 2013. Criteria for the design of corpora for monolingual lexicography. In Rufus H. Gouws, Ulrich Heid, Wolfgang Schweickard & Herber Ernst Wiegand (eds.), *Supplementary Volume Dictionaries. An International Encyclopedia of Lexicography*, 1336–1343. De Gruyter Mouton. https://doi.org/10.1515/9783110238136.1336.

Steurs, Frieda, Kris Heylen & Vincent Vandeghinste (2021). Hoe automatische vertaling de gender bias van AI verraadt. In: Wat gebeurt er in het Nederlands?. Sterck De Vreese. ISBN: 9789056158033.