Hans-Jörg Schmid

Does gender-related variation still have an effect, even when topic and (almost) everything else is controlled?

Abstract: Corpus-based studies of gender-related grammatical and lexical variation generally run the risk of underestimating the confounding effects of topic. When significant differences in the frequency of usage of certain linguistic elements and features are observed, it cannot be ruled out that they are ultimately due to gender-linked differences regarding preferred topics.

This paper presents a methodological exercise probing the question whether gender-related linguistic usage differences persist if effects of topic are neutralized. To this end, a very special corpus is exploited: the HCRC Map Task Corpus collected at the universities of Glasgow and Edinburgh, which consist of 128 dialogues revolving around the same topic. Frequency data on seven linguistic target items (*the, of, and, I, you, okay* and *mmhmm*) are collected and analyzed with regard to gender-linked differences by means of three types of regression models: negative binomial regressions, zero-inflated negative binomial regressions and mixed-effect regression models. The results of the study indicate that gender-related differences between women and men in same-gender and mixed-gender dyads can still be observed to some extent, even if the variable topic is kept constant and the functional range of the language produced is very limited.

1 Introduction

Previous research suggests that the gender of the participants involved in a conversation can affect their use of language in three different ways:

– as a function of the gender of the person speaking: female versus male (see Mulac, Bradac, and Gibbons 2001; Newman et al. 2008 for extensive surveys);
– as a function of the gender of the person addressed vis-à-vis the person speaking: same-gender talk versus mixed-gender talk (Bilous and Krauss

Hans-Jörg Schmid: Ludwig Maximilians University Munich

1988; Mulac et al. 1988; Hirschman 1994; Hancock and Rubin 2015);
- as a function of the interaction of the two: effects of the gender of the speaker that are contingent on same-gender or mixed-gender talk, or vice versa (McMillan et al. 1977; Palomares 2008).

Previous research also suggests, however, that the observable differences could simply be due women's and men's preferences regarding topics of conversation (Newman et al. 2008: 229). Women have been claimed to spend more time talking about people, past events and personal topics, while men's favourites include job-related topics, sports, politics and technology. Of course, topic choices have a strong effect on linguistic choices. For example, talk about people and past events is much more likely to contain larger numbers of proper nouns, personal and possessive pronouns, temporal and spatial adverbials as well as past tense verbs than talk about politics or cutting-edge technology. Linguistic investigations that seek to identify the effect of gender on linguistic variation are thus well advised to take the confounding effect of topic into consideration. Since topic keeps changing and drifting in casual conversation, it has turned out to be extremely difficult to control this variable in quantitative corpus studies.

It is precisely this dilemma which forms the backdrop and motivation for the present study. What is presented here is actually not much more than a methodological exercise whose key aim is to show to what extent gender-related effects on language use can still be observed if the variable TOPIC is kept constant. The characteristics of a very special dataset are exploited to reach this goal: the Human Communication Research Centre (HCRC) Map Task Corpus collected at the universities of Glasgow and Edinburgh in the 1980s (see Anderson et al. 1991 and http://groups.inf.ed.ac.uk/maptask/#top for more information)[1]. This corpus consists of transcripts of 128 dialogues, all of which had the same setup and involved the same task:

> […] two speakers sit opposite one another and each has a map which the other cannot see. One speaker – designated the Instruction Giver – has a route marked on her map; the other speaker – the Instruction Follower – has no route. The speakers are told that their goal is to reproduce the Instruction Giver's route on the Instruction Follower's map. The maps are not identical and the speakers are told this explicitly at the beginning of their

1 I would like to thank the compilers of the Map Task Corpus for sharing their material with the scientific community and Jean Carletta from the University of Edinburgh for directing me to pertinent information on the HCRC Map Task corpus website.

first session. It is, however, up to them to discover how the two maps differ (http://groups.inf.ed.ac.uk/maptask/maptask-description.html).

What makes this corpus extremely attractive for the current undertaking is that all of the 128 dialogues revolved around one topic which involves: giving directions, receiving directions and sorting out commonalities and differences between the two maps. If gender-linked differences regarding the usage frequencies of selected linguistic items can be observed in this extremely homogeneous dataset, then it seems quite certain that they are not confounded by the choice of typically feminine or masculine topics. Instead, these differences can either be correlated with the gender of the speaker, with the gender of the person addressed or with other identifiable factors such as the role of the speaker in the dialogue and the familiarity between the participants, many of which are also controlled in the dataset.

2 Research question and zero-hypothesis

The considerations sketched out so far lead to the following research question:
- Do women and men use selected words with different frequencies of occurrence if the variable TOPIC is kept constant and other variables affecting language use are also controlled?

The zero-hypothesis corresponding to this research question can be formulated as follows:
- H_0: The relative frequencies of usage of selected words (*the, of, and, I, you, okay, mmhmm*) by women and men does **not** differ in the HCRC Map Task Corpus.

3 Data, pre-processing, target variables and data retrieval

32 women and 32 men took part in the study that produced the raw material for the HCRC Map Task Corpus. All 64 persons were students at the University of Glasgow, 61 of them were native Scots. Participants were between 17 and 30 years old, with a mean age of just under 20 years. Each participant in the test served twice as Instruction Giver (GIVER) and twice as Instruction Follower (FOL-

LOWER), once talking to a person they already knew (FAMILIAR) and once to someone they were not familiar with (UNFAMILIAR). Each participant thus produced language under four different conditions: talking to a FAMILIAR and an UNFAMILIAR person in the role of GIVER and FOLLOWER. A further predictor that was controlled systematically in the study concerned eye-contact between the interlocutors. In one half of the conversations, the givers and followers could see each other (EYE.YES), in the other half there was a screen preventing eye-contact (EYE.NO). As far as the four combinations in terms of the genders of the two interlocutors are concerned, the corpus is less well balanced. In fact, as the diagrams in Figure 1 show, the number of words contributed in the different combinations varies greatly: There is much less material from mixed-gender DYADS (DYAD.MIXED) than from same-gender ones (DYAD.SAME), and, as is the case in many corpora, MALE participants contribute a considerably larger proportion to the corpus than FEMALE ones.

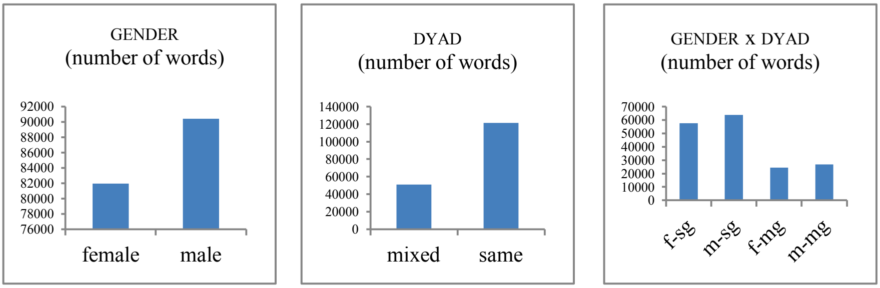

Fig. 1: Distribution of words in the HCRC Map Task Corpus across the target predictors GENDER and DYAD and their combination (absolute numbers; f-sg = female same-gender, m-sg = male same-gender, f-mg = female mixed-gender, m-mg = male same-gender)

Figure 2 provides a more detailed view of the distribution of the data by including the predictors ROLE and FAMILIARITY. The bar chart demonstrates that certain combinations of predictors are represented by a comparatively small number of observations, especially talk by FAMILIAR speakers in MIXED DYADS.

The HCRC Map Task Corpus is made available by the corpus compilers in the form of 128 files each containing one dialogue. The specific aims pursued in the present project required a substantial reprocessing of the original corpus data. The 128 original files were split in such a way that 256 files consisting of the contributions of one speaker to one conversation were created. Each of the resulting files was specified with regard to the five predictor variables: GENDER, DYAD, ROLE, EYE-CONTACT and FAMILIARITY. This revised corpus of 256 files was used in order to retrieve the frequencies of occurrence of seven linguistic target variables[2].

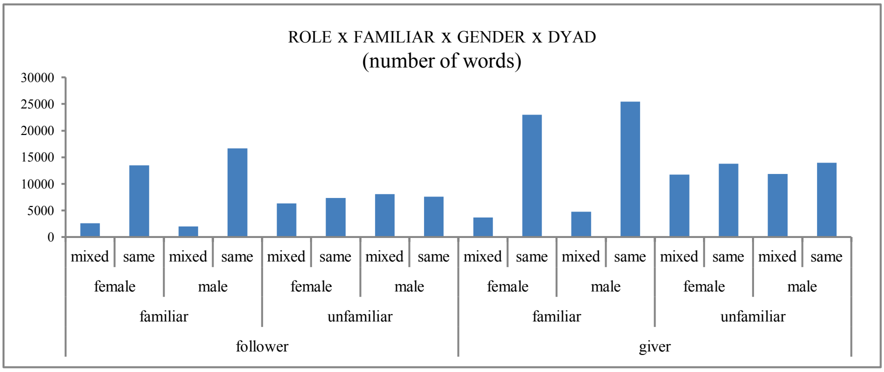

Fig. 2: Distribution of words in the dataset across target predictors (GENDER and DYAD) and the co-predictors ROLE and FAMILIARITY (absolute numbers)

The choice of these linguistic target features was complicated by the fact that the language used in the Map Task Corpus is highly functional and, as a consequence, extremely reduced regarding its lexical and grammatical complexity. This is the price that had to be paid for obtaining the thematic homogeneity which was the reason for choosing this corpus in the first place. Many linguistic target variables that have proved interesting from the point of view of gender differences such as the use of personal pronouns, past tense verbs or verbs of thinking and speaking hardly occur in the Map Task Corpus. Therefore the selection of linguistic target variables had to strike a balance between the need to collect the amount of data required for sound statistical analyses, on the one hand, and a choice of linguistic features which promised to show gender-related differences, on the other. On the basis of these inclusion criteria, the seven target variables mentioned in the null-hypothesis above were selected:

– the high-frequency grammatical items *the*, *of* and *and*;
– the deictic pronouns *I* and *you*;
– the discourse-related elements *okay* and *mmhmm*.

2 Laurence Anthony's tool *antwordprofiler* was used for this procedure (version 1200.w; see http://www.laurenceanthony.net/software/antwordprofiler/). Manual post-hoc checks were carried out using his tool *antconc*.

4 Descriptive statistics

The boxplots in Figure 3 summarize the distribution of the relative frequencies of usage of the linguistic variables in the target condition GENDER x DYAD.

The visual inspection of the boxplots in Figure 3 does not reveal any big differences with regard to the variables GENDER and DYAD and their combination. The lines indicating the medians generally do not differ much, and most of the boxes show considerable overlap. This does not give rise to the expectation that we will be seeing significant effects of the two target predictors GENDER and DYAD. The only linguistic items whose distribution could promise to yield significant gender-related differences are the discourse-related items *okay* and *mmhmm*. The dispersion of the data is generally quite high, and especially for *mmhmm* and *okay*, zero occurrences per speaker are not uncommon.

5 Inferential statistical analysis

Given the observed structure of the data, it seemed advisable to consider three different types of regression models in order to test for significant effects of the two target predictors GENDER and DYAD and the three co-predictors ROLE, EYE-CONTACT and FAMILIARITY:

- generalized linear regression models for count data which are capable of handling overdispersion, i.e. negative binomial regressions (Hilbe 2011);
- zero-inflated negative binomial regression models for count data with a large number of zero occurrences (Hilbe 2011: 370–382);
- generalized linear mixed-effect regression models taking speaker-based variation into account as random effects (Fahrmeier et al. 2013)[3].

Negative binomial models were used instead of quasi-Poisson models, since it was necessary to compare the generalized linear models to the corresponding zero-inflated models. In order to do this, the Vuong test (Hilbe 2011: 377–380) was applied. This test uses the two likelihood functions to compare negative

---

3 All calculations were carried out with the help of the software *R* (version 3.1.2). The negative binomial regression models were fitted using the `glm.nb` command from the library `MASS`, the zero-inflated models with the `zeroinfl` function from the `pscl` package, and the mixed-effects models with the `glmer` command from the package `lme4`.

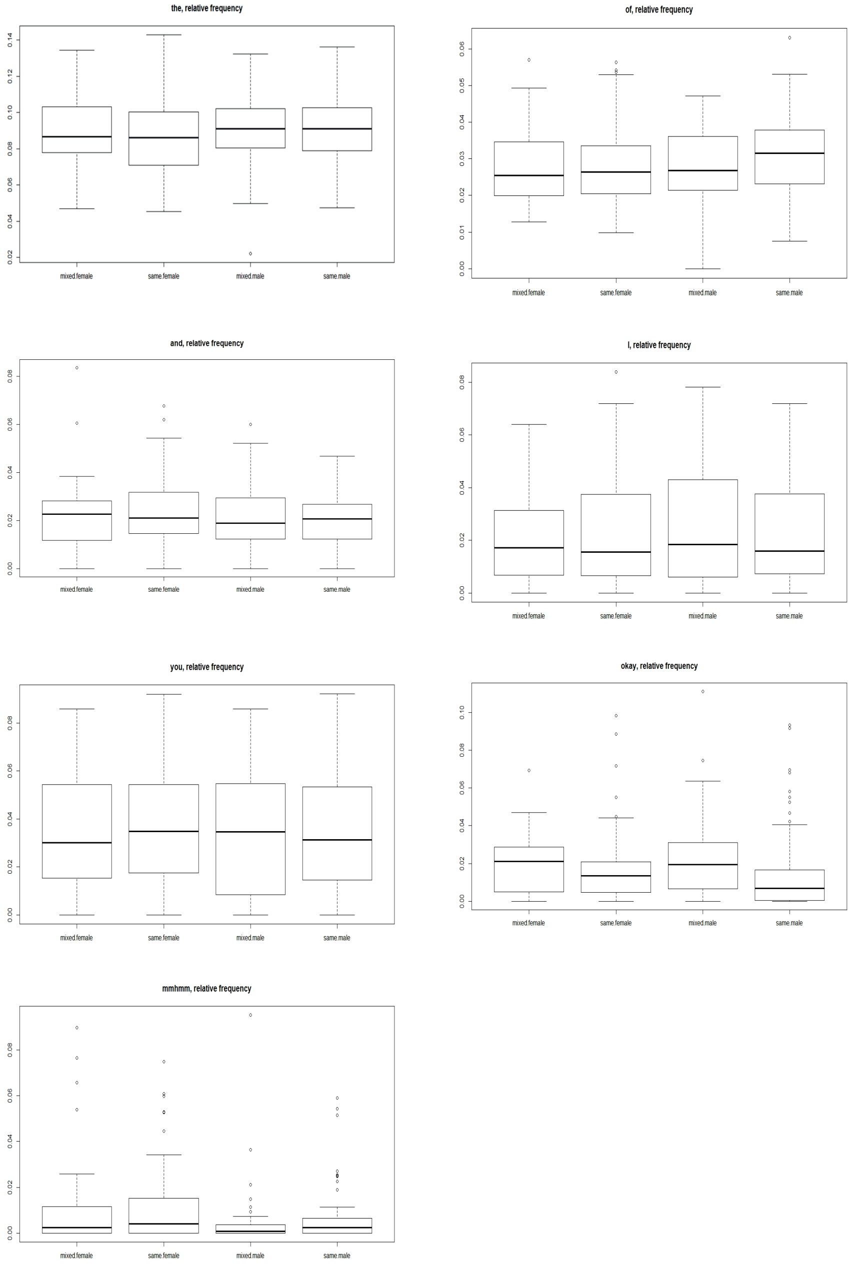

Fig. 3: Boxplots of relative frequencies of occurrence of the seven target items in the condition GENDER X DYAD

binomial models to their corresponding zero-inflated models. Since quasi-Poisson models do not have a likelihood function, they are inadequate in this situation. As negative binomial models are only capable of modeling overdispersion but not underdispersion, quasi-Poisson models were fitted first to check for overdispersion, which was confirmed by dispersion parameters well over 1 in all cases.

Tab. 1: List of predictors with significant effects on the target variables in the regression models (all effect sizes are rendered as exp(β); arrows indicate increasing/decreasing tendencies; details of all models can be found in the appendix)

predictor **linguistic variable** **n (type of model)**	GENDER	DYAD	ROLE	FAMILIARITY	EYE-CONTACT
the n = 15045 (glm.nb)					no ⇩ (p<0.1)
of n = 4891 (glm.nb)				unfamiliar 1.12 (p<0.05)	
and n = 3914 (glm.nb)	male 0.87 (p<0.01)		giver 1.70 (p<0.001)		no 0.89 (p<0.01)
I n = 3793 (glm.nb)			giver 0.23 (p<0.001)	unfamiliar 0.80 (p<0.01)	
you n = 7230 (glm.nb)			giver 3.11 (p<0.001)		
okay n = 2449 (glmer)		same ⇩ (p<0.1)	giver 0.38 (p<0.001)		no ⇧ (p<0.1)
mmhmm n = 911 (glmer)	male 0.47 (p<0.05)		giver 0.10 (p<0.001)		

The three types of models were fitted for all linguistic target variables with the aim of selecting the model suited best for the specific structure of each of them. It turned out that for the more frequent and more "grammatical" target items *the*, *of*, *and*, *I* and *you*, the random speaker effects included in the mixed-effects models did not account for any of the variance in the data. In contrast, these random effects contributed substantially to capturing variance in the models fitted for the data on *okay* and *mmhmm*. Zero-inflated models did not outperform the generalized negative binomial models for any of the linguistic target variables. It was therefore decided to accept and report the negative binomial regression models for the first five items, and the mixed-effects ones for *okay* and *mmhmm*. While interactions were generally taken into consideration, the target interaction GENDER X DYAD failed to be significant for all linguistic target variables.

The remainder of this section will provide a summary of the significant effects predicted by the models fitted for the seven target variables. This will be followed by a discussion of the findings (Section 6) and a general discussion of the results and their implications for the research question (Section 7).

Table 1 summarizes the significant effects rendered in the regression models. Effect sizes (calculated as exp(β)) and significance levels are reported. Tendencies significant at the 0.1-level are reported as well, but the effects are only indicated by arrows pointing upwards or downwards to indicate increase or decrease. More details can be gleaned from the *R* output for all models provided in the appendix.

6 Discussion

the

The negative binomial model for the definite article yields only a decreasing tendency for the NO.EYE-CONTACT condition. In contrast to the findings of previous studies (Schmid 2003;, Newman et al. 2008: 219), neither the GENDER OF THE SPEAKER nor the GENDER OF THE ADDRESSEE seem to affect the frequency of use of the definite determiner *the*. This could suggest that a considerable part of the gender-related variation found in these previous studies was at least influenced by the choices of topic and by the concomitant greater diversity of functions of the definite article. The present results indicate that if topic is held constant, the gender-related differences regarding the frequency of *the* largely disappear.

of

On the surface, the situation for the preposition *of* is quite similar. While UNFAMILIAR speakers are predicted to use *of* significantly more frequently than FAMILIAR speakers (1.12, $p < 0.05$), neither GENDER nor DYAD are listed as having significant effects. In contrast to the case of *the*, however, it is rewarding to have a closer look at the data for *of* from a gender-related perspective. Overall, the speakers in the corpus use the preposition *of* 4891 times. The very specific type of communication situation represented in the Map Task Corpus has the effect that two functions of the use of *of* strongly prevail: 3162 occurrences of the preposition occur as parts of spatial references containing the words *side, bottom, top, left, edge, right, middle, end, corner, level, site, outside, centre, point* and *tip*. Another 903 occurrences are parts of hedges or vague complex quantifiers using the nouns *sort* and *kind,* and *couple* and *bit* respectively. What is remarkable about this functional distinction is that the predictor GENDER affects these two usage-types in fundamentally different ways. This is demonstrated by regression models that were fitted separately for the different portions of the data[4]. As shown in Table 2, the predictions made by these models differ substantially.

Tab. 2: List of predictors credited with significant effects on different uses of the preposition *of*

Predictor **linguistic variable** **n (type of model)**	**GENDER**	**DYAD**	**ROLE**	**FAMILIARITY**	**EYE-CONTACT**
of used in frequent spatial references n = 3162 (glm.nb)	male ⇧ (p<0.1)			unfamiliar 1.13 (p<0.05)	
of used in hedges and vague quantifiers n = 903 (glmer)	male 0.68 (p<0.05)		giver 1.38 (p<0.05)		no ⇧ (<0.1)
other uses of *of* n = 826 (glm.nb)	male 1.46 (p<0.001)	same 1.27 (p<0.05)	giver 0.80 (p<0.05)		no 0.81 (p<0.05)

The negative binomial model for the general spatial-reference uses of *of* yields a significant increasing effect for dialogues between UNFAMILIAR participants (1.13, p<0.05). There is only a tendency for *of* to be used more frequently by men. In contrast, MALE GENDER turns out to be a significant predictor with a decreasing, rather than increasing effect on the frequency of *of* used in hedges and vague

4 For the spatial references, a mixed-effects model was not required because the random speaker effect did not capture any of the variation, while for the hedging use of *sort of* etc., the opposite was the case.

quantifiers (0.68, $p<0.05$), alongside the increasing effect of the role of GIVER (1.38, $p<0.05$). The strongest gender-linked effects are found in the negative binomial regression for the remaining 826 attestations in the corpus, which predicts a strong increase for MALE GENDER speakers (1.46, $p<0.001$) and SAME-GENDER DYADS (1.27, $p<0.05$). In addition, the model yields a decreasing effect for GIVERS (0.80, $p<0.05$) and for NO EYE-CONTACT situations (0.81, $p<0.05$). The manual inspection of this portion of the data reveals that the strongest gender difference is found for very precise spatial references using fractions (*three quarters of, a third of, two thirds of* and *half of*) and cardinal or intermediate directions (*west of, east of, northwest of* etc.). The men in the corpus use these types of references more than 3.5 times more often than the women, which confirms earlier findings on the overuse of spatial references by men (Mulac and Lundell 1986: 89). In addition, the present results corroborate a number of stereotypes frequently voiced especially in the older language-and-gender literature (cf. e.g. Lakoff 1975): Women are likely to use hedges and vague language more frequently than men, while men are more likely to produce very precise spatial references, especially when talking to other men.

and

For the target variable *and*, the negative binomial model reveals significant effects for the predictors ROLE, EYE-CONTACT and also GENDER (cf. Table 1). The frequency of *and* is predicted to rise by a factor of 1.70 ($p<0.001$) for the role of GIVERS as opposed to FOLLOWERS, and to drop in the NO EYE-CONTACT condition (0.89, $p<0.05$). In addition, the model predicts a drop by a factor of 0.87 for MALE as opposed to FEMALE speakers. While it seems rather difficult to interpret these findings, a closer inspection of the data indicates that the effect of GENDER can be attributed at least partly to two frequent and functionally similar types of sequences used by GIVERS while instructing FOLLOWERS where to go on the map: the complex continuative *and then* (e.g. *down towards the east and then back up again*) and sequences of spatial adverbs and the conjunction *and*, most frequently *right/left/up/down and* (e.g. *you turn right and go straight across*). These usage types account for 1133 and 447 occurrences respectively and thus for about 40% of the total of 3914 uses of *and*. A negative binomial model for this part of the data, which is also reported in the appendix, predicts a strong and significant decreasing effect for MALE GENDER (0.69, $p<0.001$). The corresponding model for the remaining 2440 uses of *and* does not include a significant effect for GENDER. This means that it seems legitimate to conclude that it is first and foremost the targeted subset which accounts for the effects of GENDER on the variable *and*. This specific usage of *and* as a general-purpose continuative typical of spontaneous speech (cf. Biber et al. 1999: 81–83) can be related to other features claimed to be overrepresented in the speech of women which signal high speaker involvement and conversational commitment (Tannen 1990).

I

A considerable part of the variation of the variable *I* is explained by the dominant variable ROLE, with GIVERS being predicted to be significantly less likely to use this pronoun than FOLLOWERS (0.23, $p<0.001$). A second predictor with significant effect is FAMILIARITY (0.80, $p<0.01$). While findings by Schmid (2003) and Newman et al. (2008: 219) suggest that women use the first-person singular pronoun more frequently than men, this does not seem to be the case in conversations of this functionally restricted type and when topic is controlled.

you

The use of the target variable *you* is dominated by a single equally strong and predictable variable: a massive increase by a factor of 3.11 ($p<0.001$) associated with the ROLE of GIVER. None of the other predictors comes close to achieving significant effects on the distribution of *you*. And, just for the record, for *you*, Newman et al. (2008: 220) observe a significant increasing effect for males, while Schmid (2003) observed a preponderance of *you* in female talk. This contradiction is not resolved by the analysis of the special dataset investigated in the present study.

okay

The mixed-effects model for the discourse marker *okay* also reveals a very strong effect of the predictor ROLE, viz. a decrease by a factor of 0.38 ($p<0.001$) for GIVERS. This is not surprising, since FOLLOWERS are much more likely to signal uptake than GIVERS. In addition, the model predicts tendencies for the variables EYE-CONTACT and DYAD. A preponderance of the use of *okay* in the speech of men, which is suggested by the analysis of the British National Corpus (BNC) reported in Schmid (2003), is not confirmed.

mmhmm

The mixed-effects model for the backchannel item transcribed as *mmhmm* in the corpus predicts a significant effect of GENDER. According to this model, MALES are significantly less likely (0.47, $p<0.05$) to produce this signal of active listenership than FEMALES. This result concurs with existing findings that men are more reluctant to show involvement and to contribute actively to the smooth flow of conversation (e.g. Zimmermann and West 1975). In addition, ROLE is again included in the model as a very strong predictor (0.10, p<0.001).

To summarize, the GENDER of the speaker is predicted to have significant effects on the frequencies of usage of the backchannel item *mmhmm*, on continuative uses of the conjunction *and* and on uses of the preposition *of* in the hedging constructions *sort of* and *kind of,* in the vague complex quantifiers *couple of* and *bit of* and in precise spatial references using fractions (*a third of* etc.) and cardinal and intermediate directions (*north of* etc.). The gender of addressees vis-à-vis the gender of the speaker, i.e. the variable DYAD, was shown to have significant effects on the frequencies of occurrence of the remaining varied uses of the preposition *of*. Limited as these results are, they still mean that the zero-hypothesis formulated in Section 2 must be rejected. Gender-related differences in frequencies of usage can indeed be observed for certain linguistic elements even if the variable TOPIC is kept constant.

The nature of the findings generally indicates that gender-specific language use seems to be dominant in the field of discourse-related elements: The variables *mmhmm*, continuative *and* as well as hedges and vague quantifiers including the preposition *of* turned out to be affected by GENDER and/or DYAD, while the frequency of items such as *the, I* and *you*, whose use is more strongly determined by grammatical and immediate pragmatic needs, seems to be immune to the influence of these factors, at least if TOPIC is as strictly controlled as in the present dataset. It is possible that the use of the discourse-related elements leaves more room for individual speaker habits and routines. This assumption would also be supported by the finding that the random speaker effects were mainly relevant for modelling these types of target variables.

7 General discussion and conclusion

The language use and the choices of linguistic variants by given speakers in given situations are known to be subject to a wide range of factors: user-related variables such as the REGIONAL and SOCIAL BACKGROUND, EDUCATION, GENDER, AGE and ETHNICITY of the speaker, on the one hand, and use-related variables such as SETTING, PLACE, TIME, MEDIUM, PARTICIPANTS (and their user-related traits) as well as SUBJECT-MATTER and TOPIC, on the other. The user-related variable GENDER OF SPEAKER, the use-related variable GENDER OF ADDRESSEE and possible interactions between them served as target predictors of the present study. What was special about it is that an extraordinarily large number of potential confounds was controlled: The medium was spontaneous spoken speech throughout; all speakers were students of approximately the same age; almost all of them were native Scots and spoke Scottish English; all conversations were dialogues taking place under controlled conditions regarding the roles of the two participants and the familiarity and eye-contact between them. Plus, all conversations shared the same topic. The merit for all this of course goes to the compilers of the HCRC Map Task Corpus.

The regression models presented above indicate that the variables GENDER OF SPEAKER and GENDER OF ADDRESSEE have significant effects on the frequencies of occurrence of four of the seven target items. The research question posed in the title of this paper can therefore be answered with a cautious "yes, to some extent gender-related variation continues to have an effect on language use, even when topic and virtually everything else is controlled". It is hoped that this insight is of use in further studies on language and gender and sparks off further investigations.

Three pre-final caveats are called for: As pointed out above, the language used in the course of solving the map task is very special in terms of its limited functional diversity and reduced linguistic complexity. It is not unlikely that a greater degree of gender-related variation would be observed if TOPIC was controlled in a less strict way, so that speakers remained free to exploit the full lexical and grammatical resources of the spoken medium. Second, the material making up the HCRC Map Task Corpus was collected in the early 1980s when gender roles and identities, both linguistic and otherwise, were different from what they are today. One would hope that a replication of this study with material elicited under identical conditions today would yield different results. And third, while the present findings could possibly have far-reaching implications for studies of language and gender, to go into these implications was beyond the scope of the present contribution.

This leaves me with one final question I want to address to the dedicatee of this volume: What on earth, in the context of the present paper, could be "the meaning of variation" (Geeraerts and Kristiansen 2014)?

Acknowledgement

I am greatly indebted to Franziska Günther, Daphné Kerremans and Gill Woodman for their comments on an earlier version of this paper, and to Alexander Bauer from the Statistical Consulting Unit at Ludwig Maximilians University

Munich directed by Helmut Küchenhoff for his expert advice on the statistics used in this paper.

Appendix

Tab. 3: Negative binomial regression for the target variable *the*

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	−2.4289195	0.0390445	−62.209	<2e-16	***
gendermale	0.0364854	0.0260763	1.399	0.1618	
dyadsame	0.0004823	0.0283781	0.017	0.9864	
rolegiver	−0.0197464	0.0248840	−0.794	0.4275	
eyeno	−0.0465606	0.0260839	−1.785	0.0743	.
familiarunfamiliar	0.0382661	0.0259835	1.473	0.1408	

Tab. 4: Negative binomial regression for the target variable *of*

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	−3.740793	0.071983	−51.968	<2e-16	***
gendermale	0.078108	0.047544	1.643	0.1004	
dyadsame	0.083790	0.05185	1.616	0.1061	
rolegiver	0.005032	0.045278	0.111	0.9115	
eyeno	0.062443	0.047582	1.312	0.1894	
familiarunfamiliar	0.117902	0.047208	1.473	0.0125	*

Tab. 5: Negative binomial regression for the target variable *and*

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	−4.03461	0.08652	−46.63	<2e-16	***
gendermale	−0.13964	0.0575	−2.428	0.0152	*
dyadsame	0.01205	0.06249	0.193	0.8471	
rolegiver	0.53262	0.05601	9.51	<2e-16	***

	Estimate	Std. Error	z value	Pr(>|z|)	
eyeno	−0.11715	0.05756	−2.035	0.0418	*
familiarunfamiliar	0.07602	0.05725	1.328	0.1842	

Tab. 6: Negative binomial regression for the target variable *I*

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	−3.17289	0.11122	−28.527	< 2e-16	***
gendermale	0.11893	0.07722	1.54	0.12354	
dyadsame	−0.12504	0.08337	−1.5	0.13368	
rolegiver	−1.48285	0.07353	−20.165	< 2e-16	***
eyeno	0.105	0.07727	1.359	0.17418	
familiarunfamiliar	−0.21924	0.07672	−2.858	0.00426	**

Tab. 7: Negative binomial regression for the target variable *you*

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	−4.01174	0.077743	−51.602	<2e-16	***
gendermale	−0.0614	0.050535	−1.215	0.224	
dyadsame	0.026062	0.054983	0.474	0.636	
rolegiver	1.135047	0.050558	22.45	<2e-16	***
eyeno	0.003375	0.050595	0.067	0.947	
familiarunfamiliar	−0.00762	0.050275	−0.152	0.879	

Tab. 8: Mixed-effects regression for the target variable *okay*

Random effects:

Groups	Name	Variance	Std.Dev.
id.sp	**(Intercept)**	0.3409	0.5839
Residual		0.5142	0.7171

Number of obs: 256, groups: id.sp, 64

Fixed effects:

	Estimate	Std. Error	z value	Pr(>|z|)

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	−3.8538	0.3163	−12.186	< 2e-16	***
gendermale	−0.2026	0.2576	−0.787	0.4315	
dyadsame	−0.3621	0.1992	−1.818	0.0691	.
rolegiver	−0.9647	0.1407	−6.859	6.95E-12	***
eyeno	0.4554	0.2611	1.745	0.0811	.
familiarunfamiliar	−0.1313	0.1505	−0.873	0.3829	

Tab. 9: Mixed-effects regression for the target variable *mmhmm*

Random effects:

Groups	Name	Variance	Std.Dev.
id.sp	**(Intercept)**	0.6832	0.8265
Residual		0.6965	0.8345

Number of obs: 256, groups: id.sp, 64

Fixed effects:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	−4.22432	0.36821	−11.473	<2e-16	***
gendermale	−0.75671	0.32106	−2.357	0.0184	*
dyadsame	0.15061	0.23381	0.644	0.5195	
rolegiver	−2.28771	0.17111	−13.37	<2e-16	***
eyeno	−0.32986	0.32195	−1.025	0.3056	
familiarunfamiliar	−0.08016	0.17105	−0.469	0.6393	

Tab. 10: Negative binomial regression for the target variable *side/bottom/top* etc. *of*

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	−4.10042	0.09244	−44.356	<2e-16	***
gendermale	0.10665	0.06154	1.733	0.0831	.
dyadsame	0.01863	0.06672	0.279	0.78	
rolegiver	−0.01571	0.05854	−0.268	0.7885	
eyeno	0.05444	0.06156	0.884	0.3766	
familiarunfamiliar	0.12661	0.06114	2.071	0.0384	*

Tab. 11: Mixed-effects regression for the target variable *sort/kind/couple/bit of*

Random effects:

Groups	Name	Variance	Std.Dev.
id.sp	**(Intercept)**	0.0760	0.2757
Residual		0.6533	0.8082

Number of obs: 256, groups: id.sp, 64

Fixed effects:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	−5.81885	0.27043	−21.517	<2e-16	***
gendermale	−0.38501	0.18828	−2.045	0.0409	*
dyadsame	0.18588	0.18859	0.986	0.3243	
rolegiver	0.31925	0.15469	2.064	0.039	*
eyeno	0.34392	0.18746	1.835	0.0666	.
familiarunfamiliar	0.06137	0.16121	0.381	0.7035	

Tab. 12: Negative binomial regression for remaining 826 occurrences of *of*

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	−5.54344	0.15213	−36.438	< 2e-16	***
gendermale	0.38152	0.09944	3.837	0.000125	***
dyadsame	0.24198	0.10976	2.205	0.027481	*
rolegiver	−0.22328	0.09283	−2.405	0.016165	*
eyeno	−0.20587	0.09823	−2.096	0.036099	*
familiarunfamiliar	0.05659	0.09735	0.581	0.561029	

Tab. 13: Negative binomial regression for *and then* and *right/left/up/down and*

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	−5.08935	0.143783	−35.396	< 2e-16	***
gendermale	−0.36419	0.0957	−3.806	0.000142	***
dyadsame	0.007006	0.103521	0.068	0.946046	
rolegiver	0.793976	0.094503	8.402	< 2e-16	***

	Estimate	Std. Error	z value	Pr(>|z|)	
eyeno	−0.19461	0.09573	−2.033	0.042057	*
familiarunfamiliar	0.198483	0.095154	2.086	0.036987	*

Tab. 14: Negative binomial regression for remaining uses of *and*

Coefficients:					
	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**	
(Intercept)	-4.48682	0.101574	-44.173	< 2e-16	***
gendermale	-0.05841	0.06702	-0.872	0.383	
dyadsame	0.002675	0.066984	0.04	0.968	
rolegiver	0.017901	0.073068	0.245	0.806	
eyeno	0.376237	0.065092	5.78	7.47E-09	***
familiarunfamiliar	0.000516	0.066748	0.008	0.994	

References

Anderson, Anne H., Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry S. Thompson & Regina Weinert. 1991. The HCRC Map Task Corpus. *Language and Speech* 34(4), 4351–4366.

Biber, Douglas, Susan Conrad, Geoffrey Leech, Stig Johannson & Edward Finegan. 1999. *Longman grammar of spoken and written English*. London: Longman.

Bilous, Frances R. & Robert M. Krauss. 1988. Dominance and accommodation in the conversational behaviors of same- and mixed-gender dyads. *Language and Communication* 8(3-4). 183–194.

Fahrmeier, Ludwig, Thomas Kneib, Stefan Lang & Brian Marx. 2013. *Regression: Models, methods and applications*. Berlin & Heidelberg: Springer.

Geeraerts, Dirk & Gitte Kristiansen. 2014. Cognitive linguistics and language variation. In Jeannette Littlemore & John Taylor (eds.), *The Bloomsbury companion to cognitive linguistics*, 202–217. London & New York: Bloomsbury.

Hancock, Adrienne B. & Benjamin A. Rubin. 2015. Influence of communication partner's gender on language. *Journal of Language and Social Psychology* 34(1). 46–64.

Hilbe, Joseph M. 2011. *Negative binomial regression*, 2nd edn. Cambridge: Cambridge University Press.

Hirschmann, Lynette. 1994. Female-male differences in conversational interaction. *Language in Society* 23(3). 427–442.

Lakoff, Robin T. 1975. *Language and woman's place*. New York, NY: Oxford University Press.

McMillan, Julie R., A. Kay Clifton, Diane McGrath & Wanda S. Gale. 1977. Women's language: Uncertainty or interpersonal sensitivity and emotionality? *Sex Roles* 3(6). 545–559.
Mulac, Anthony & Torborg L. Lundell. 1986. Linguistic contributors to the gender-linked language effect. *Journal of Language & Social Psychology* 5(2). 81–101.
Mulac, Anthony, James J. Bradac & Pamela Gibbons. 2001. Empirical support for the gender-asculture hypothesis: an intercultural analysis of male/female language differences. *Human Communication Research* 27(1). 121–152.
Mulac, Anthony, John W. Wiemann, Sally J. Widenmann & Tony W. Gibson. 1988. Male/female language differences and effects in same-sex and mixed-sex dyads: The gender-linked language effect. *Communication Monographs* 55(4). 315–335.
Newman, Matthew L., Carla J. Groom, Lori D. Handelman & James W. Pennebaker. 2008. Gender differences in language use: An analysis of 14.000 text samples. *Discourse Processes* 45(3). 211–236.
Palomares, Nicolas, A. 2008. Explaining gender-based language use: effects of gender identity salience on references to emotion and tentative language in intra- and intergroup contexts. *Human Communication Research* 34. 263–286.
Schmid, Hans-Jörg. 2003. Do women and men really live in different cultures? Evidence from the BNC. In Andrew Wilson, Paul Rayson & Tony McEnery (eds.), *Corpus linguistics by the Lune: A Festschrift for Geoffrey Leech*, 185–221. Frankfurt: Peter Lang.
Tannen, Deborah. 1990. *You just don't understand: Women and men in conversation*. New York, NY: Morrow.
Zimmermann, Don H. & Candace West. 1975. Sex roles, interruptions and silences in conversation. In Barrie Thorne & Nancy Henley (eds.), *Language and sex: Difference and dominance*, 105–129. Rowley, MA: Newbury House.