Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2018”

His Fault, but Ihre Wünschen [Her Desires]: A Corpus-based Study on Language and Gender

Linnik I. (julietlinn@icloud.com), Saint Petersburg State University, Philological faculty, Department of General Linguistics, Saint Petersburg, Russia

This study presents a comparative and cross-linguistic (German and English) analysis of how men and women are referred to in the web. For this investigation we are seeking for associations that are thriving in chosen languages. Our focus is on the most frequent nouns determined by pronouns *his/her* and *ihr*/sein** in enTenTen13 and deTenTen13 (Sketch Engine) respectively. By examining the nouns we look at how women and men are linguistically represented and whether it depends on a language. Some of the nouns were subsequently chosen for a more careful study encompassing framework analysis and a greater attention to adjectives that can define the nouns in different contexts.

Keywords: language and gender, interactional sociolinguistics, corpus linguistics

HIS FAULT, IHRE WÜNSCHEN [ЕГО ВИНА, ЕЕ ЖЕЛАНИЯ]: ИССЛЕДОВАНИЕ СВЯЗИ ГЕНДЕРА И ЯЗЫКА МЕТОДАМИ КОРПУСНОЙ ЛИНГВИСТИКИ

Линник Ю.В. (julietlinn@icloud.com), Санкт-Петербургский государственный университет, филологический факультет, кафедра общего языкознания, Санкт-Петербург

1 Introduction

Any language can be analyzed from various points of view, and one of them is the approach of sociolinguistics, which is concerned with linguistic variation and its social significance. In this particular branch of linguistics, the concept of dividing people into different social categories – gender, age, ethnicity, social networks, etc. – plays a sufficient role, because the use of language may vary depending on the social group. Those slightly different speech types are seen to contribute to social meanings. (Eckert and Rickford 2001) Therefore, the object of sociolinguistic studies is the language distinction between members of different social groups.

Two main approaches of sociolinguistics are mentioned usually: variationist paradigm (William Labov and others) and interactional sociolinguistics (John Gumperz, Dell Hymes and others). The last one would be of a special interest for our investigation as it encompasses many subtopics, which mainly discuss relations between language, the way of thinking and, therefore, acting of people while performing communication, so that code-switching framework would be revealed. "The theory brings together several major sets of resources: linguistics and discourse analysis, (…) Goffmanian and conversation analysis, (…), ethnography, (…) [and finally,] Gumperz adds the vital notions of conversational *inferencing* and *contextualisation*". (Rampton 2017)

Technological breakthrough of XX century led to the formation of new scientific approaches, which provide researchers with faster implementation of the objectives. Quantitative and qualitative methods of Corpus Linguistics are used with "intrinsic explanatory purpose", so that it would be easier to engage linguistics and the language itself with theories of other Humanities' branches. (Mahlberg, 2015) According to the book "Corpus Linguistics and the Description of English" by Hans Lindquist, corpus needs to "be marked up with information about speakers and writers" (Lindquist 2013). However, it is also possible to make an investigation about links between gender and language in a slightly more subtle way - without special tagging, but with the analysis of the most frequent collocations of certain gender-marked words. (Lindquist 2013) We found it interesting to carry out a small research about adjectives and nouns, which collocate with pronouns *his*/*her* and *sein*/ihr** in English and German Corpora respectively. That analysis might reveal us peculiar language associations, which unintentionally concern gender and therefore capture social reality in itself.

In our research we adopted the approach used by Anna Čermakova and Lenka Farova in their article "His eyes narrowed — her eyes downcast: contrastive corpus-stylistic analysis of female and male writing" (2017). They have investigated Czech and English fiction with a view to variation in male and female writers' styles, which resulted in a closer analysis of appearance description. However, for our investigation the gender of a writer at this point is not that relevant as we are seeking for associations that are thriving in the language itself - partly because of history, partly as a result of strict grammar rules and structures both in German and English.

Therefore, our presumption consists of two hypotheses:

(1) Most of results for both genders in English are probably a combination of possessive pronouns with nouns describing of appearance (body parts, head and face) or nominating family members, as it is required by English grammar. In German either the construction with a reflexive verb taking a Dative pronoun would be preferred in the same context, e.g. *Ich wasche* ***mir*** *die Hände*, but *I wash* ***my*** *hands*;
(2) There might be a marked difference between English and German as those two languages belong to different cultures and have different structures.

2 Method and Material

As the main source of texts for analysis, Sketch Engine was chosen. It contains both English and German Corpora, which are built under the same principles and therefore seem to be comparable

to each other. The German web corpus deTenTen13 consists of 16.5 billion tokens, while the English one enTenTen13 includes a bigger amount of words – 19.6 billion. According to the documentation from the web site, all of the corpora include mostly non-fiction texts and a relatively small amount of fiction. EnTenTen13 includes more texts of British origin, rather than of Australian, Canadian or American, but still it is a notable amount (see the table below).

DeTenTen13 is composed of texts not only from the German domain (14.3 billion tokens), but also from domains of the countries where German is used as an official language (2.1 billion). As it was mentioned above, we are interested in general language patterns, and in this particular study we will not be concerned with English and/or German language varieties.

As Čermakova and Farova did in Czech-English comparative research, for the first step we are going to analyze the 50 most frequent nouns occurring immediately after *his* and *her* and decide, which of them seem to be more "masculine" or "feminine". For the second step, we will analyze more closely some of the most frequent nouns in both languages. Statistical significance is going to be tested with chi square distribution.

3 Typical "Female" and "Male" Descriptions

3.1 'His/Her + Noun','ihr*/sein*+Noun'

We created a list of 100 most frequent words for each group, counted ipm for each line so that results would be more accurate, and cleared out the list from accidental, but predictable lines like situation of adnominal possessive relation in English, e.g. ***her father's*** *X*, or article and flexion variation in German because of grammatical cases, e.g. *sein Vater*, *seines Vaters*, *seinem Vater*.(Haspelmath 2008) Other obstacle was grammatical disambiguation of pronouns *ihr* in the beginning of a sentence and *Ihr*; for that we had to study concordances manually. Finally, we got the list of the 50 most frequent nouns occurring next to the possessive pronouns in the context and divided them into five semantic groups:

- Family and Friends
- Career (which also includes words like *money*, *salary*, *colleagues*, etc.)
- Time and Life, or the way of life differently (which includes words like *life*, *death*, *opinion*, *love*, *aim*, *wish*, etc.)
- Body description
- Other (the least significant group, which includes only two words for all languages and genders – *book*, *car*)

Afterwards we counted the general ipm for each of the five groups; you may see the result below.

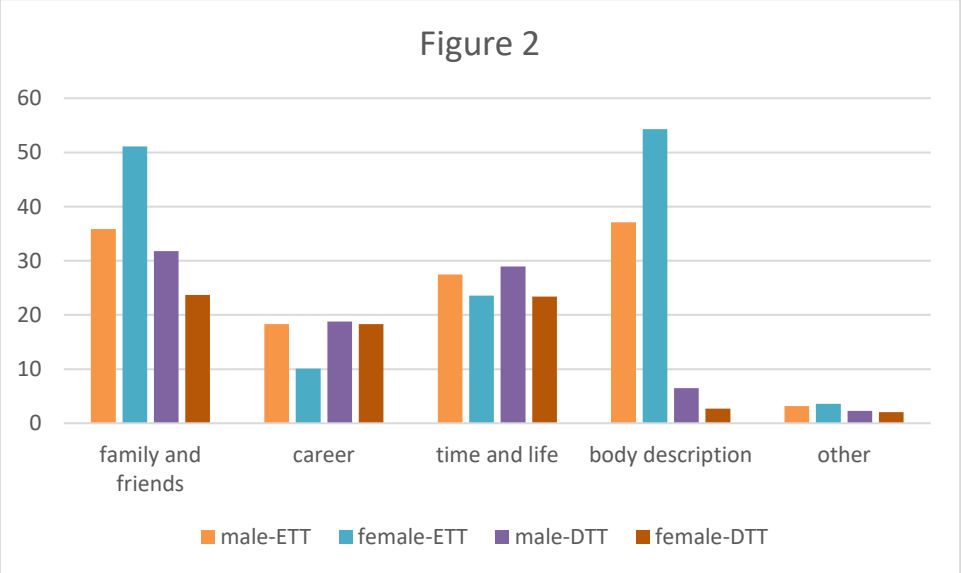

Figure 1. Semantic groups' distribution within gender and language (ipm sum)

As it is seen from the diagram, nouns for family members and friends are more often used in the group female-ETT. In this group we meet more frequent and varied usage of phrases about children, e.g. we may note *her child*, *her baby* and *her kids*(cf. *his children* and *sein/ihr Kind*). Moreover, in both male-DTT and female-DTT, but not in their English counterparts, (*ihre/seine*) *Gäste* turned to be one of the most frequent combinations. In addition, we should mention phrase *ihr Engagement*, which is not frequent (as *her engagement*) even in the female-ETT. Surprisingly, overall frequency of female-DTT is twice less than frequency of female-ETT and has the smallest result overall. For all groups the most often mentioned relatives are *husband* and *wife*. However, the phrase *seine Frau* has the frequency of 7 ipm, and *ihr Mann* has the frequency of only 3 ipm, which is considerably small in comparison with English results (10,7 ipm for female-ETT and 9,8 ipm for male-ETT). Perhaps, female family life in Germany (and other German-speaking countries) is less exposed in publicity. We may also assume that women are just less discussed in texts from geTenTen13 than men or than both genders in enTenTen13.

The group Career seems to have almost equal distribution for male-DTT, male-ETT, and female-DTT. In German there were more words related to career and work among the 50 most frequent nouns than in English. It is explained by the fact that German tends to use English loanwords and original words at the more or less same frequency: *das Werk* and *die Arbeit*, *die Karriere* and *die Laufbahn*. However, that does not affect the general statistics. What is more interesting, phrases *sein Team* and *his team* have rather high frequencies - 1,8 and 2,5 ipm respectively, but *her team* and *ihr team* could not enter the final list. From the analysis of our concordance, it seems that in most cases the possessive pronouns in combination with words like *team*, *company*, *business* imply that the person is the leader of a group or the owner/head of a company, but not its part or employee. In the female-ETT, the words from this group were poorly represented and had the lowest frequency. The female-ETT includes only three most general words of this kind: *work*, *career* and *job*, while all other groups (male-ETT, …, …) have at least words *salary*, *order*, *money*, *position*, *office*. As a result, we may say that in both corpora there are more male-leaders than female.

Like the previous group, Time and Life is more or less uniform, and none of the sub-groups showed drastically high or considerably low results. However, the difference is seen in the further analysis. For instance, the word *death/Tod* is used more often with "male" pronouns *his*, *ihr*, and in the female-DTT it is not frequent at all. This probably may be a mold of a tragic statistics from gender and longevity area, which shows that men tend to die sooner than women. (Regan and Partridge 2013) In the female-DTT we may find plenty of words referring to wishes: *ihre Wünschen*, *ihre Bedürfnisse*, *ihre Können*, which is not characteristic of the other groups. For both male-DTT and male-ETT we may observe rather frequent use of words *choice/Wahl*, *fault/Schuld*, and *sein Mannschaft* only for German.

Contrary to our expectations, the German corpus seems to contain less body descriptions in comparison to the English one. Probably, that is just the indicator of a lesser amount of fiction texts in deTenTen13 as well as particularities of German grammar. However, there is a considerable amount of female-body description in enTenTen13. It is not hard to notice that some of the body parts are frequent only for female-ETT: *her fingers*, *her feet*, *her back*, *her lips*. The other remarkable fact is that the word *chest* tends to be more "masculine" than "feminine", as in the female-ETT its frequency is 10 times lower than in the male-ETT (0,01 and 0,1 ipm respectively).

3.2 How Do Men and Women Look Like: Eyes and Hair

According to the statistics from the previous part, the most frequent group for enTenTen13 is body descriptions. The words *Augen* and *Haar* also fall within the top-ten words in the list of frequencies for deTenTen. Therefore, we decided to study whether there is a difference between descriptions of hair and eyes in German and English.

There is a noticeable overlap in the lists of the most frequent adjectives preceded by pronouns *his/her*, *ihr*/sein** and followed immediately by words *hair* or *eyes* afterwards. Most of them are apparently colors and shades, and this tendency does not substantially vary between groups. However, it is possible to consider them as one of the characteristics of a person.

Feature / Group	light/ warm	dark	cold	tired	big	small
female-ETT	0,068	0,116	0,000	0,030	0,038	0,007
male-ETT	0,044	0,176	0,006	0,062	0,025	0,022
female-DTT	0,051	0,053	0,000	0,025	0,072	0,003
male-DTT	0,033	0,107	0,013	0,043	0,039	0,009

Table 1. Correlation between gender and the description of eyes (ipm sum)

From the table above it is clearly seen that there is almost no difference between languages, but rather between "masculine" and "feminine" descriptions. Women's eyes in both languages tend to be bright, big, wide and expressive, while men's eyes seem to be depicted as cold, tired, small and heavy. No special adjective for 'coldness' of the eyes is found for women in both corpora, while for men we found *stahlblauen*, *eisblauen* and just *cold*.

English corpus seems to contain more fiction, therefore it includes more peculiar and diverse adjectives for eyes. Below you may see a table with adjectives that tend to be more "feminine" or "masculine" according to the search results gained from enTenTen13.

Female-ETT	**ipm**	**Male-ETT**	**ipm**
her violet eyes	0,0085	his burning eyes	0,015
her emerald eyes	0,0085	his yellow eyes	0,0117
her crescent eyes	0,0081	his eagle eyes	0,0098
her sparkling eyes	0,0066	his bloodshot eyes	0,0097
her pretty eyes	0,0060	his sharp eyes	0,0090
her wet eyes	0,0051	his glittering eyes	0,007
her great eyes	0,0051	his sleepy eyes	0,0068
her expressive eyes	0,0051	his piercing eyes	0,0068
her lovely eyes	0,0050	his keen eyes	0,0067
her wonderful eyes	0,0045	his heavy eyes	0,0060
her sharp eyes	0,0044	his cold eyes	0,0060

Table 2. Adjectives for eyes that tend to be more "masculine" or "feminine" in enTenTen13

From those instances, it could be noticed that the English language uses more 'lovely' and 'wonderful' descriptions for women's eyes, and 'heavy' and 'piercing' for men.

Slightly different situation is observed in the descriptions of hair. In English there are still more depictive adjectives, because of the same reasons mentioned above. Nevertheless at least three fascinating tendencies are present. Firstly, unnatural colors are used more frequently to describe hair in female-ETT and male-DTT, e.g. *her pink hair*, *seine grüne Haare*. Secondly, the connotation of wilting and ageing is more typical for men, but not for women in both languages, e.g. *seine ergrauten Haare, his thinning hair* are more frequent than the corresponding combinations with *her*, *ihr*. Thirdly, only for men's hair exist plenty of frequent synonyms for shagginess in both languages, and in contrast, some special words are found as implied [by the context] hairstyle only for women, e.g. his – *messy, shaggy, matted, tousled*; sein* - *zerzausten, verwuschelten*; her – *loose, loosened*; ihr* - *glatten, gesammelten*. However, we should bear in mind that those two web-corpora do not contain enough fiction and, therefore, they might provide us with statistically imprecise results.

3.3 Average Life

Returning slightly back and seeking through the main list again, it would not be surprising to observe that the word *life/Leben* has the highest ipm in deTenTen13 and the second highest in enTenTen13 for both genders. Consequently, we could not ignore the question, which adjectives are usually used for description of men's and women's lives. Overall, the results seem to be more or less similar, because more than half of the combinations 'possessive pronoun + adjective + *life*' for each gender include autobiographical clichés: *his former life*, *her past life*, *ihr zukünftiges Leben*, *sein ganzes Leben*. We decided to focus our attention on the rest of examples, which do not imply the expression of time.

Feature / Group	Public life and work	Everyday life, home life, family
female-ETT	0,21	0,31
male-ETT	0,38	0,16
female-DTT	0,09	0,27
male-DTT	0,19	0,08

Table 3. Characteristics of lives (ipm sum)

We divided all results into two categories. It is clearly seen above that a man's life is combined more frequently with adjectives that are related to such topics as working life, political life, and public life. The "routine" adjectives like *day-to-day*, *everyday*, *alltäglichen* etc. seem to be more frequent in the female-ETT and female-DTT groups, which may probably mirror the social position of a woman who usually stays at home all day long, and her social role is to care about home life. Moreover, in the female-ETT there is sought a significant instance *her married life,* which should also be interpreted as a language picture of the reality described above. However, significally frequent instances with *married* or *verheiratet* cannot be found in any other group (neither any male, nor female-DTT), which may indicate that either German society (or German web) is slightly more open-minded and stereotype-free, or most likely there is a lack of needed texts in deTenTen13.

4 Conclusions

By and large, we may conclude that there is a significant difference between German and English concept of gender (at $p<0,025$), but we cannot be sure about the absolutely similar composition and equal distribution of texts within the two corpora used in our study. Moreover, the results related to the female-DTT showed the lowest overall frequency in all of the groups from the first step (see part 3.1), which might mean that women are less discussed in deTenTen13 (or in the German web in general). On the contrary, the female-ETT confirms all hypotheses we made in the introduction and represents all typical stereotypes: family, home life and body descriptions are twice more frequent in combination with *her* than with *his*, while both male-DTT and male-ETT have higher results related to working and public life. The word *death/Tod* is used more often with male pronouns, and in the female-DTT it is not frequent at all. This probably may be a mold of a tragic statistics from gender and longevity area, which shows that men tend to die sooner than women.

As a result of the German grammatical limitations, we found drastically small number of results with *sein*/ihr*+body parts* (much less than expected). It seems also that enTenTen13 includes more fiction texts than deTenTen13. Therefore, there are more body descriptions in the English subcorpus and more fruitful adjectives. Nevertheless, some common tendencies for genders without depending on languages are sought. Women's eyes in both languages tend to be bright, big, wide and expressive, while men's eyes seem to be depicted as cold, tired, small and heavy. In addition, in the texts of our corpora men tend to have shaggy hair much more often than women, while women tend to have hair done.

These very general conclusions seem to confirm some gender stereotypes, which are generally the same for both languages. However, the deeper understanding of the concept needs a more careful study on some other samples and other corpora, which would have more fiction and/or some extracts from social networks: Facebook, Twitter, etc. Anyhow, stereotypes are still sought and, therefore, exist. “Gender differences in language phenomenon are not accidental; they have the profound social root”. (Dong 2014) It is the result of different social roles, duties and rights of a deep historical origin.

5 References

Čermáková A., Fárová L. (2017), His eyes narrowed — her eyes downcast: contrastive corpus-stylistic analysis of female and male writing, Linguistica Pragensia, Vol. 28, pp 7-34.

Eckert P., Rickford J.R. (2001), Style and Sociolinguistic Variation, Cambridge University Press, Cambridge.

Haspelmath M. (2008), Alienable vs. inalienable possessive constructions, Syntactic Universals and Usage Frequency, Leipzig Spring School on Linguistic Diversity, March 2008.

Jinyu D. (2014), Study on Gender Differences in Language Under the Sociolinguistics, Canadian Social Science, Vol. 10, pp. 92-96.

Lindquist H. (2013), Corpus linguistics and the description of English, Edinburgh University Press, Edinburgh, UK.

Mahlberg M. (2015), Literary style and literary texts, The Cambridge Handbook of English Corpus Linguistics, Cambridge University Press, Cambridge, pp 346-361.

Rampton B. (2017), Interactional Sociolinguistics, available at: www.kcl.ac.uk

Regan J.C., Partridge L. (2013), Gender and longevity: Why do men die earlier than women? Comparative and experimental evidence, Best Practice & Research: Clinical Endocrinology & Metabolism, Vol. 27, pp. 467-479.