Exercise 1 – Conditional Frequency Distributions

1a) Table

Pronouns		He	Him	She	Her
Genres	**News**	642	93	77	121
	Religion	206	94	12	8
	Government	169	26	1	3
	Fiction	1308	382	425	413
	romance	1068	340	728	680

1b) Answer in words what you see. How does gender vary with the genres?

As it becomes obvious from the table above – produced in exercise 1a – the masculine pronouns are in general more frequent than the feminine pronouns. The masculine pronouns outnumber the feminine ones in every single genre being looked at except for romance (male vs. female: 1408 each). This might be the case because romance supposedly seems to be a genre which is being read/consumed by women more often than men and therefore there might be more women in prominent roles here. In all other genres, the masculine pronouns outweigh the feminine ones extremely. There is even the incident that the feminine pronoun “she” only appears a single time in the “government” genre and the pronoun “her” only appears three times in that same genre, which makes it an almost exclusively male genre.

When it comes to comparing the frequency distribution of “he” and “she” over the different genres, the masculine pronoun outnumbers the feminine one in every single genre and also with a huge difference in numbers. Considering the frequency distribution of “him” and “her” across the genres, the feminine pronoun appears more often than the masculine counterpart in three out of five categories (news, fiction and romance). One reason for that might be that women in those genres might be the object of the action (“her”) more often than in other genres, and usually a male actor is the subject being referred to as “he”.

One last observation to mention is that “her” seems to be more frequent compared to “she”, while with respect to the masculine forms “he” dominates over “him”. Here, there seems to be a different dynamics in the use of the two forms of the genders.

1c) Table and Numbers

Form	Male pronouns	Female pronouns
Nominative forms (he, she)	9548	2860
Objective forms (him, her)	2619	3036

Relative Frequency 'Her' from ‘She’ or ‘Her’: 0.5149253731343284

Relative Frequency 'Him' from ‘He’ or ‘Him’: 0.21525437659242214

1d) Table

Pronouns	Personal pronouns	Possessive pronouns	Total
She	2860	0	**2860**
He	9546	0	**9546**
Her	1107	1929	**3036**
Him	2619	0	**2619**
His	0	6957	**6957**
Hers	0	0	**0**
Total	**16132**	**8886**	**51965**

1e) Corrected Numbers on Pronouns

Relative Frequency 'Her' 0.2790521804890345

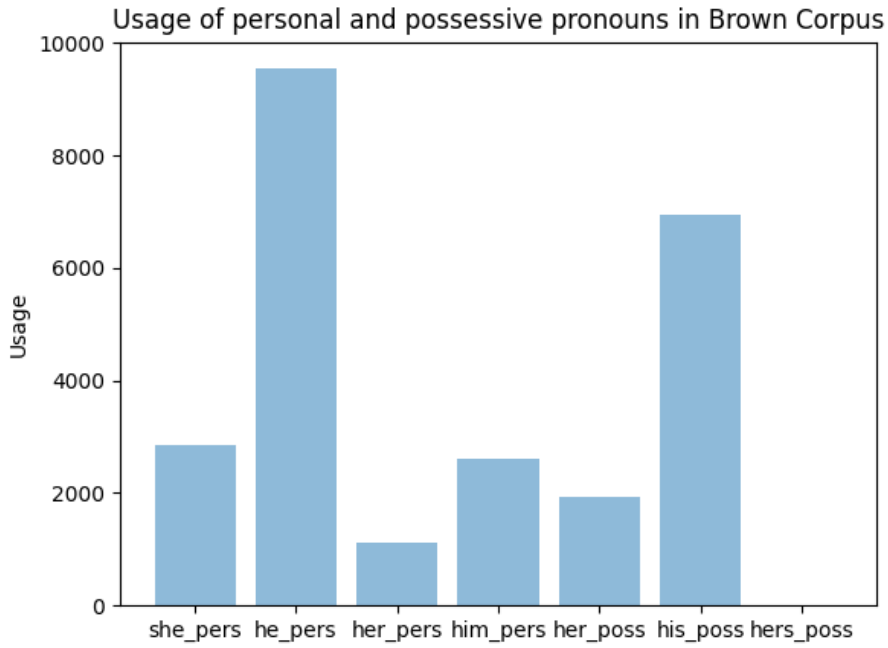
Relative Frequency 'She' 0.7209478195109654

Relative Frequency 'He' 0.7847102342786683

Relative Frequency 'Him' 0.21528976572133168

As we can see now, the percentages of “her” compared to “him” are now quite even, as well as “she” compared to “He”. Therefore, it was important to filter the instances of “her” being a possessive pronoun instead of a personal pronoun – now the values are much more comparable to each other and serve research purposes more properly.

1f) Bar chart with data from d



1g) Essay

As it becomes clear from the findings of the research above, the masculine pronouns appear more frequently in language than their feminine counterparts. This might be because the masculine pronouns are used more often when the speaker doesn't know the gender of the person they are talking about or when referring to an unknown entity. In English, one can use "one" or "they" to make a neutral reference, but still the male pronouns are used as well. Furthermore, it could also be that male persons are featured more often in written texts and therefore their pronouns are used more frequently than female pronouns.

Based on the findings described in the previous exercises, a number of consequences arises for the development of language technology, especially considering the data used in training. Firstly, language technology should be tested and developed not only on a huge amount of data, but on a variety of different genres of texts as well. Like this, one can prevent or at least try to prevent the development of a technology which, e.g., discriminates against women because it wasn't trained properly on female pronouns etc. Moreover, it is insurmountable that the data used for the development is recent and not too old. Language is constantly changing and therefore the models should be trained on the most recent data possible. Further, the training of the models should continue while they are already being used, which ensures that the models will keep up with language change. Concluding this, and drawing on the knowledge about the Brown corpus, the corpora used for training the language models should be constantly updated and redesigned, so that the genres incorporated are genres which still exist in the current culture and not outdated ones.