MOHAMMED ALTAMINI – SERGEI GNITIEV – ANNA ISMAIL JAROUR – ŻANETA KATONA – ALA EDDIN KHELIFA – ABDOUSS MOHAMED – DIANA OKTAVIA – MARIIA POPOVA – HERLAND AKBARI PUTRA – IBTISSEM SMARI – VINCENT J. VAN HEUVEN

Pannon Egyetem, Veszprém, Hungary
mm_tamimi83@yahoo.com; raymax43@yandex.ru; ana_2020_2020@hotmail.com; przybyla.zaneta@gmail.com; khelifaalaeddin@gmail.com; mohamed.abdouss@usmba.ac.ma misssdiana@ymail.com; mariiapopova.izh@gmail.com; esotericingenium@gmail.com; smariibtissem@yahoo.fr; v.j.j.p.van.heuven@hum.leidenuniv.nl

Altamini–Gnitiev–Jarour–Katona–Khelifa–Mohamed–Oktavia–Popova–Putra–Smari–van Heuven: Gender effects on writing style in British English?
Alkalmazott Nyelvtudomány, XVII. évfolyam, 2017/2. szám
doi:http://dx.doi.org/10.18460/ANY.2017.2.006

Gender effects on writing style in British English?

A tanulmány az angol nyelvű írásokban érvényesülő gender-hatást vizsgálja. Arra a kérdésre keressük a választ, hogy milyen mértékben hasonlóak az angol nyelven író női és férfi szerzők által használt szavak és mondatszerkezetek. 36 brit angol nyelvváltozatban készült, női és férfi szerzőktől származó, a közelmúltban megjelent cikket választottunk, amelyek párba állítva egyazon témáról fogalmaznak meg véleményt. Munkánk során elemeztük a szövegekben előforduló szókincs kiterjedését, a szavak terjedelmét (leütések száma), a mondatok terjedelmét (mondatonkénti szavak száma), a mondatok szerkezetét (tagmondatok vagy állítmányi szerepben használt igék száma), valamint az érzelmeket kifejező és egyéb írásjelek előfordulását. Az eredmények több szövegjellemző esetén a női és férfi írásokban egyezésre utalnak, amely a téma hatására hívja fel a figyelmet. Várakozásaink ellenére a női írók több alárendelő mondatszerkezetet alkalmaznak az írásokban mint a férfiak, azonban nemre vonatkozó egyéb szignifikáns eltérés nem azonosítható. Következtetésünk szerint az idegen nyelv oktatásában nemek szerint eltérő angol nyelvű írásfejlesztés nem szükséges.

1. Introduction

This article is a summary of work done by a group of master students in the Department of Applied and Hungarian Linguistics at the University of Pannonia. In the one-week all-day course on Corpus Linguistics students were required, first of all, to study a text book on Corpus Linguistics (McEnnery–Wilson, 2001). The chapters in the book were presented, one a day, by the instructor in a series of PowerPoint presentations after which the students did a selection of the exercises listed at the end of each chapter. Their solutions were sent in the evening as e-mail attachments to the instructor who commented on the results at the start of the next lecture. In order to prevent the course from becoming an exercise in treading the beaten path, it was decided to complement the textbook work with hands-on experience in collecting textual data according to some systematic design,

enriching the textual data with linguistic information, and using the assembled materials as an empirical basis to test specific hypotheses on the use of language by specific social groups. We explicitly did not query existing corpora, since that would preclude the students from collecting and enriching their own data.[1] Data manipulation and basic statistical analysis were included as part of the teaching goals.

The student group was highly heterogeneous in terms of linguistic background. Students were from a diversity of nationalities with differences in native languages to match. Two spoke Russian, four spoke different varieties of Arabic (Moroccan, Tunisian, Palestinian), two were native speakers of Indonesian, and one was bilingual in Polish and Hungarian. Given this diversity of language backgrounds and since the language of communication in the lectures was English, it was decided to set up a small-scale research project on English language use and approach it with Corpus Linguistic methods. The purpose of the project was not to query an existing language corpus but to build a (necessarily small) corpus from scratch, and enrich it with the lexical and syntactic information needed to answer specific research questions – a situation which would approximate the students' future professional work realistically.

The starting point for the project was the often heard claim that men and women typically use different language words and language structures. Almost a century ago the renowned Danish linguist Otto Jespersen (1922) made the following rather sweeping statements with respect to the different use of language between male and female speakers (and presumably authors).[2] With respect to gender differences in vocabulary Jespersen (1922: 248) says:

> (…) the vocabulary of a woman as a rule is much less extensive than that of a man. Women move preferably in the central field of language, avoiding everything that is out of the way or bizarre, while men will often either coin new words or expressions or take up old-fashioned ones, if by that means they are enabled, or think they are enabled, to find a more adequate or precise expression for their thoughts.

In corpus-linguistic terms this claim translates to the hypotheses that (i) women use fewer different words, i.e., their token-to-type ratio should be larger than that of men, and (ii) women tend to use high-frequency words whereas men tend to also use words with low token frequencies in the language at large, and even invent

words that have never been used before (typically compounds assembled from existing words).

Jespersen (1922: 251) makes the further claim that women use parataxis rather than hypotaxis, while hypotaxis (deemed a more complex and intellectually more challenging sentence structure) would rather be characteristic of male language use:

> If we compare long periods as constructed by men and by women, we shall in the former find many more instances of intricate or involute structures with clause within clause, a relative clause in the middle of a conditional clause or vice versa, with subordination and sub-subordination, while the typical form of long feminine periods is that of co-ordination, one sentence or clause being added to another on the same plane and the gradation between the respective ideas being marked not grammatically, but emotionally, by stress and intonation, and in writing by underlining. In learned terminology we may say that men are fond of hypotaxis and women of parataxis.

This claim would lead us to expect women to typically use short sentences while men express their thoughts in longer and more complicated sentence structures typically containing multiple finite verbs, each being the pivotal word of a separate, embedded clause. Moreover, the claim that women take recourse to prosodic means (or the written expression thereof) to express their thoughts emotionally would lead us to expect a higher prevalence of non-neutral punctuation marks in female writings, i.e. question mark and exclamation marks, in contradistinction to male texts which would use the more neutral sentence-final punctuation mark, i.e. the full stop.

Interestingly, Jespersen (1922: 253) believed that the use of simple words and sentence structures allows women to express their thoughts more quickly than men:

> The superior readiness of speech of women is a concomitant of the fact that their vocabulary is smaller and more central than that of men.

Moreover, Jespersen does not claim that the female use of language is inferior to that of men on average. He points out that the difference in language use between the two sexes is found in the extremes (1922: 253):

> (…) it may serve as a sort of consolation to the other sex that there are a much greater number of men than of women who cannot put two words together intelligibly, who stutter and stammer and hesitate, and are unable to find suitable expressions for the simplest thought.[3] Between these two extremes the woman moves with a sure and supple tongue which is ever ready to find words and to pronounce them in a clear and intelligible manner.

We decided to put these (and similar, more recently formulated) claims as to differences in language use between men and women to the test. For this purpose we collected a relatively concise corpus of written language use by male and female authors who regularly publish their opinions through articles on the internet. If it is true, as Jespersen argued, that the differences between the sexes should not be sought in the center of the distribution but in the extremes, the results should all the more clearly show differences between male and female authors.

An author's vocabulary and the complexity of the sentences needed to express one's ideas are obviously related to the topic the writing deals with. A political essay will generally contain different words than, say, an article about a fashion show. In order to make sure that our statistics should not be contaminated by the choice of typically male versus female subjects chosen by the authors, we set as a constraint that the authors who contributed to our corpus should be matched pairwise in terms of topic. So, for instance, for every male author who wrote an essay on the presidential elections in the United States of America in the autumn of 2016, we would pair this with a female author writing about the same topic.

The project is interesting not only from the point of view of pure research, i.e. for its potential contribution to our knowledge of differences in language between socially defined categories of speakers and writers, such as gender-related differences. We also do the research in order to assess the extent to which it would be necessary to develop different course materials for the teaching of English writing skills to male and female students – whether native or non-native. For instance, if we were to find that female authors use shorter and more frequent words than men, and use shorter sentences with fewer embeddings, then presumably it would be desirable to mirror these differences in the goals of English writing courses for foreign students. This, of course, would complicate the teaching of English as a foreign language to a considerable extent, and impose a burden on the foreign language instructors we would preferably avoid.

2. Hypotheses

In light of the background presented in the previous section we developed a number of hypotheses at the word and at the sentence level, which we will list and motivate in the present section.

At the word level:

H1 Women will use shorter words than men – all else being equal. Here, words are practically defined as any string of letters bounded by spaces or punctuation marks (including the hyphen). If it is true that women use simpler words with higher token frequencies than men, we should find mean word length (expressed as number of letters in the word) to be shorter in female than in male texts. We make this prediction on the strength of one of Zipf's laws, which states that words in the lexicon are shorter (and have more different meanings) as they are used more often (Strauss et al. 2007, Zipf 1932, 1935). Using this negative correlation between word length and word frequency avoids the necessity of importing and analysing the token frequencies of the words in our corpus – which is a project for the future.

H2 The token-to-type ratio (TTR) in female texts will be larger than in male writings. This hypothesis is a technical expression of the claim that men employ a greater variety of words, i.e. use the same word less often, than women. This hypothesis also follows from the assumption that women prefer words from the central (i.e. most frequently used) part of the vocabulary.

H3 Women use more function words than men. This hypothesis follows from the observation that men want to express themselves in less ambiguous terms, and therefore choose to avoid referring to entities in the non-linguistic context by means of deictic elements such as personal pronouns. It was shown in the British National Corpus that female authors use personal pronouns more often than their male colleagues (Argamon et al. 2003). The present hypothesis generalizes this idea to the total set of function words.

At the sentence level:

H4 The length (expressed as number of words) in sentences written by female authors will be shorter than those of male authors. This is the crudest way of testing the idea that women express themselves by linguistically simple structures: the shorter the sentence, the easier it will be to understand. The sentence will be practically defined as any string of words bounded by

punctuation marks that define the terminal boundary of a sentence, i.e. the full stop, the (semi-)colon, the exclamation mark and the question mark.

H5 Sentences written by men, even if there is no difference in number of words, will be more complex than those written by women. Although sentence length and complexity are correlated in practice, length and complexity are independent parameters in principle. Therefore we study two complexity measures in addition to length. Sentence complexity can be expressed as the number of subclauses per sentence. The more subclauses, the more complex the sentence. Given that each subclause must contain a finite verb, counting the number of finite verbs per sentence would be a useful index of sentence complexity. We may also reason that subordinate (embedded) clauses are more difficult to process than coordinated clauses, so that the former contribute ore to sentence complexity than the latter. To differentiate between the two complexity measures, we defined a second sentence complexity measure more specifically as the number of embedded clauses (hypotaxis rather than parataxis) per sentence. In order to do so we need to count the number of subordinating pronouns (relative pronouns) and paratactic conjunctions per sentence.

H6 Women use non-neutral sentence-final boundary marks (i.e. question mark and exclamation mark) more often than men do. This hypothesis transfers Jespersen's observation (see above) that women tend to express their feelings and emotions by speech melody rather than by verbal means as men do, from the spoken to the written modality of language use.

3. Procedures

Eighteen pairs of texts were located on the internet, each text being a recent column or editorial-style discussion of political, educational or cultural developments written by a British English author. The two texts making up a pair had to deal with the same subject. One text had to be written by a male author, its counterpart by a female author. Appendix 1 lists the titles and topics of the pairs of male and female-authored texts we used. Texts were downloaded from internet websites using copy-paste procedures and saved as plain text files after tables, graphs and pictures had been removed. Then, using the AntConc concordance software (Anthony, 2012, 2013), a word list with token frequencies was generated for each text and saved as a Microsoft Excel workbook. The length (in letters) was

computed in Excel using the string length function. Mean word length (weighted for word frequency) was computed as well as the mean token frequency of the words.[4]

The normalized text files were then uploaded to the CLAWS part-of-speech tagger on the server at Lancaster University. This PoS-tagger was developed to facilitate the tagging of the British National Corpus and is claimed to perform with 97-97% accuracy (Garside–Smith, 1997). We used the fairly restricted C5 tagset, which recognizes just over 60 different parts of speech.[5] The output that is returned by the tagging service provides sequence numbers for the sentences within the text and for words within sentences, so that it is easy to compute the length of each sentence in terms of number of words, and from that the mean and standard deviation of the sentence length. For each of the 36 texts we then determined the number of content words by adding up all the occurrences of tags that define content words.[6] The proportion of content words is computed by dividing the total number of content words by all the words in the text. Again using the CLAWS5 tags, we also determined the number of finite verbs per text. Mean sentence complexity is then conveniently expressed as the number of finite verbs (= clauses) divided by the number of sentences. This measure covers both coordinated (parataxis) and subordinated (hypotaxis) clauses per sentence. Alternatively, we counted the number of subordinating conjunctions and relative pronouns (tags CJS, CJT, DTQ) which is a good estimate of the number of number of subordinate clauses. The sentence complexity is then expressed as the number of subclauses per sentence (hypotaxis only). As a last exercise we counted the number of question marks and exclamation marks and computed from this the proportion of non-neutral (or "emotional") closures.

4. Results

The dataset contained 36 texts, 18 written by male authors and 18 more by female authors. The total number of words amounted to just under 50,000 (47,461), almost equally distributed between the sexes (24,019 for the male sample and 23,542 for the female authors). The lengths of individual texts varied considerably, also within male-female author pairs. Table 1 presents a summary of the results. The table lists for each male-female author pair the mean value found for each of the textual parameters defined in section 3. The table also specifies the difference between the genders (Δ = female − male), the t-statistic computed for

correlated samples and the probability of obtaining such a difference in means by chance.

Table 1. Summary of results. Mean values for eight textual parameters broken down by gender of author. Means are based on 18 texts by 18 different authors. The difference (Δ female − male), the t-value for correlated samples (df = 17) and p-value are given. Significant differences ($p < .05$, two-tailed) between genders are indicated by *

Parameter	Female	Male	Δ	t(17)	p
1. Word length (letters)	4.9425	4.9442	−0.00179	−0.030	.976
2. Token / type ratio	2.3998	2.4589	−0.05907	−0.476	.640
3. Content words (%)	45.4965	45.0852	0.41129	0.130	.898
4. Sentence length (words)	20.3230	19.2633	1.05973	0.937	.362
5. Finite verbs / sentence	1.9621	1.8543	0.10780	0.730	.476
6. Subordinate clauses / sent	0.7130	0.5420	0.17104	1.665	.114
7. Ratio parataxis / hypotaxis	2.1157	2.7099	−0.59421	−2.252	.038*
8. Emotional punctuation (%)	3.4651	6.3286	−2.86342	−1.435	.170

The results show that there are hardly any differences between the male and female authors. There is no significant difference in either mean word length or in the token-to-type ratio. The mean word length is 5 letters and each word is used between 2.4 and 2.5 times per text.

Sentences written by female authors are marginally longer than those of men (20 vs. 19 words equally long), and women use slightly more content words. Both differences are counter to the prediction but the effects are, again, totally insignificant. Also, counter to the predictions, the female authors use slightly more complicated sentences, as is evidenced by the number of finite verbs per sentence. If only simplex sentences were used, each sentence should contain precisely one finite verb. In our samples the number of finites is almost two per sentence. A minority of finite verbs occur in subordinate (embedded) clauses, more so for female than for male authors. Again, none of the gender differences here reaches statistical significance. Interestingly, however, the prevalence of subordination over coordination is stronger for women than for men. Although this finding, too, runs counter to predictions that can be derived from the literature, it is the only difference in the data that reaches statistical significance. Finally, the results suggest that men use more emotional punctuation marks than women, but this time the difference, which runs against the prediction, is not significant.

Table 2 lists the correlation coefficients found between the male and female textual parameters. If it is true that the particular topic imposes restrictions on the vocabulary (technical terms and terminology) or use of complex sentences, we should find significant correlations for at least some of these textual properties.

Since we entertain the hypothesis that the topic of writing should affect all of the textual parameters at issue, the correlation coefficients can be one-tail tested for significance. Table 2 shows that the textual properties tend to be correlated between male and female authors. Weak and insignificant correlations are found for five of the parameters. However, for the three remaining parameters Word length, Percentage of content words, and Number of finite verbs per sentence (in descending order of magnitude) we observe moderate and statistically significant correlations – which finding is in line with our expectation and which shows that our decision to apply the t-test for correlated samples was correct.

Table 2. Pearson correlation coefficients for values on textual parameters by male and female authors writing on the same topics. N = 18. Significant correlations ($p < .05$, one- tailed) are denoted by *

Parameter	r	p
1. Word length (letters)	.588	.005*
2. Token / type ratio	−.112	.330
3. Content words (%)	.528	.012*
4. Sentence length (words)	−.149	.277
5. Finite verbs / sentence	.425	.040*
6. Subordinate clauses / sent	.302	.112
7. Ratio parataxis / hypotaxis	.283	.127
8. Emotional punctuation (%)	.254	.155

5. Conclusion and discussion

In section 2 we developed six hypotheses which can be tested on the results described above. Let recapitulate the hypotheses and examine what conclusions can be drawn for each of them.

We hypothesized that female authors would use shorter words than male authors, and related to that, that they would use more common words with higher frequencies. The results do not support this hypothesis. If male and female authors write about the same topics, there is no difference in word length. The same goes for the second hypothesis, which said that women would use fewer different words

to express their thoughts. We found, counter to this hypothesis, that the number of different words per unit length of text, expressed as the token-to-type ratio, was the same across genders.

By the same token, our results do not support the view that men tend to express themselves more clearly by using, relatively speaking, more content words than women. Moreover, we found no difference in sentence length between male and female authors, even though it was predicted that men would use longer and more complicated sentences. With respect to complexity, two measures were defined. One was the number of clauses per sentence, irrespective of the coordinate versus subordinate status of the clauses. On aggregate, we found no difference in the number of clauses (defined as the number of finite verbs) per sentence as a function of gender of author. Testing the more specific hypothesis that men use more embedded (i.e. subordinated) clauses per sentence than women, we found in fact the opposite: the female authors in our sample tended to use subordination relatively more often than their male colleagues.

Jespersen (1922) conceded that the number of complex sentences might not differ between men and women but suggested that women use coordinated subclauses (parataxis) while men would rather use (the cognitively more demanding) subordinated clauses (hypotaxis). This hypothesis was tested by examining the ratio of paratactic over hypotactic structures. The result runs clearly counter to Jespersen's claim: our female authors show a significantly greater prevalence of hypotaxis than the men.

Our last hypothesis took its cue from Jespersen's (and others') idea that women would rely more on prosody to express their emotion than men. In spite of this idea we found no indication that our female authors end their written sentences more often with non-neutral (emotional) punctuation marks. In fact, a higher percentage of the sentences produced by the male authors ended in either a question mark or an exclamation mark than the sentences of the female authors – but the difference was insignificant.

It should be realized, of course, that Jespersen's claims about gender-related differences in language use primarily pertain to spoken language. One may legitimately raise the issue whether the claims made for spoken language can reasonable be tested on data obtained from written language use. We would argue that Jespersen's ideas can be extended to the use of written language. His claim is not that men in general are better language users than women but only that the

quality of expression through language is more varied in men than in women. Women's language proficiency is rather tightly clustered around the population mean while men's proficiency may vary between extremely poor and extremely good (see the quotation from Jespersen, 1922: 253 in the introduction part of this article). We believe that journalists and authors of newspaper editorials belong to the upper branch of language users. If Jespersen's ideas would be correct, the prediction follows from this that the differences between male and female language users should be even more visible in written language than in spoken language. However, obviously, this is not what we found in our results.

In sum, our exercise contradicts the traditional and still widespread stereotypical idea that men are the better and more sophisticated writers. Differences observed in the past were quite probably caused by the choice of the topics women typically talked and wrote about.[7] Our data show that, indeed, the choice of topic affects the vocabulary and sentence complexity employed by the author but at the same time, differences due to gender turn out to disappear when men and women write about the same things. These conclusions confirm suggestions found in earlier research. For instance, phonological and pragmatic differences between male and female language usage have been reported for spoken English (Kunsmann 2000) but are not expected in formal written texts, in which phonological and conversational hints to the gender of the author would be severely reduced (Simkins-Bullock & Wildman 1991).

On the strength of these conclusions, finally, we see no need for formulating separate goals and developing different teaching materials for English writing courses for male versus female foreign students.

Notes

1. Students were advised to read the study by Argamon et al. (2003) in which the British National Corpus (BNC) was queried in terms of gender-related differences between texts in a range of genres written for an unseen readership.
2. In this article we do not systematically distinguish between social gender (as the individual's self-perceived or projected image of masculinity versus femininity) and biological gender (or: sex). For a discussion of the distinction and its consequences for the use of language and speech we refer to Biemans (2000). The attribution of gender to the authors whose writings we collected was entirely based on information available on the internet (first name, photograph).
3. Jespersen (1922: 254) explains the differences between male and female language use from an evolutionary perspective. Men have little use for language since they are traditionally engaged in activities such as war and hunting, with little engagement of linguistic interaction and where silence may be of the essence. Women, on the other hand, raise the children, work the land and prepare

food, i.e., less strenuous activities which leave ample room for social interaction through language. More recently the gender-related differences have been explained in rather more ethological terms such that women are primarily interested in getting the message across from sender to receiver with the simplest possible means so as to reduce the risk of misunderstanding, while men try to impress their peers (and the opposite sex) with unusual words and involved sentence structures very much the same way as is observed among animals. In the animal kingdom the males of the species typically display more exuberant and variegated behavior (in outward appearance and repertory of sounds and movements) than the females, who are typically smaller and subdued (see e.g. Haan & Van Heuven 1999 and references therein).

4. Each participant located two pairs of texts on the internet. As part of the seminar each student performed all the manipulations and computations needed to produce the statistics for his or her texts. The data were then aggregated and analyzed by inferential statistics. Here students worked in pairs, where each pair was instructed to test one of the six hypotheses formulated and produce a joint report on the exercise.
5. A listing of the C5 tagset can be obtained from http://ucrel.lancs.ac.uk/claws5tags.html.
6. This was done in MS Excel by executing a find-replace action for a disjunction of all content word tags (using wild card conventions). The set of content word is found by the following disjunction: AJ? (= AJ0, AJC, AJS), AV0, NN? (= NN0, NN1, NN2), NP0, VV? (= VVB, VVD, VVG, VVI, VVN, VVZ). The number of replacements was output by Excel and stored for later statistical analysis.
7. In support of this view, there is an extensive literature on the different topics typically dealt with by men and women (e.g. Aries & Johnson 1983, Biemans 2000, Tannen 1990). The difference in favored topics in written language was recently demonstrated in a statistical study of book reviews published in The New York Times between 2000 and 2015 (Piper & So 2016). In the reviews of books authored by women the words that came up most frequently were a different set than those used in reviews of male-authored books. The following quotation illustrates the gender effect convincingly:

> Book reviewers are three or four times more likely to use words like “husband,” “marriage,” and “mother” to describe books written by women (…), and nearly twice as likely to use words like “love,” “beauty,” and “sex.” Conversely, reviewers are twice as likely to use words like “president” and “leader,” as well as “argument” and “theory,” to describe books written by men. The results are almost *too good* in their confirmation of gender stereotypes. *New York Times* book reviews overwhelmingly suggest that women tend to write about domestic issues and affairs of the heart, while men thrive in writing about “serious” issues such as politics. It’s not that women don’t write about politics or men don’t write about feelings and families. It’s just that there is a very strong likelihood that if you open the pages of the *Sunday Book Review*, you will be jettisoned back into a linguistic world that more nearly resembles our Victorian ancestors.

References

Anthony, L. (2012) AntConc (Version 3.3.5) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/.

Anthony, L. (2013) Developing AntConc for a new generation of corpus linguists. In: Hardie, Andrew and Love, Robbie (eds.) *Proceedings of the Corpus Linguistics Conference (CL 2013)*, Lancaster University, UK, 14–16.

Argamon, S., Koppel, M., Fine, J., Shimoni, A. R. (2003) Gender, genre, and writing style in formal written texts. *Text – An interdisciplinary Journal for the Study of Discourse*, 23, 321–346.

Aries, E. J., Johnson, F. L. (1983) Close friendship in adulthood: Conversational content between same-sex friends. *Sex Roles*, 9, 1183–1196.

Biemans, M. (2000) *Gender variation and voice quality* (LOT dissertation series 38). Utrecht: LOT. Retrieved from https://www.lotpublications.nl/Documents/38_fulltext.pdf.

Garside, R., Smith, N. (1997) A hybrid grammatical tagger: CLAWS4. In: Garside, Roger, Leech, Geoffrey and McEnery, Anthony (eds.) *Corpus annotation: Linguistic information from computer text corpora*. London: Longman, 102–121.

Haan, J., Heuven, **V. J. van** (1999) Male vs. female pitch range in Dutch questions. *Proceedings of the 14th International Congress of Phonetic Science*s, San Francisco, 158–1584.

Jespersen, O. (1922) *Language. Its nature, development and origin*. London: George, Allen & Unwin.

Kunsmann, P. (2000) Gender, status and power in discourse behavior of men and women. *Linguistik online*, 5(1). Retrieved from https://bop.unibe.ch/linguistik-online/article/view/1017. (DOI: http://dx.doi.org/10.13092/lo.5.1017).

McEnery, T., Wilson, **A.** (2001) *Corpus Linguistics. An Introduction*. Edinburgh: Edinburgh University Press

Piper, A., So, R. J. (2016) Women write about family, men write about war. *The New Republic*, 8 April 2016. Retrieved from https://newrepublic.com/article/132531/women-write-family-men-write-war

Simkins-Bullock, J. A., Wildman, **B. G.** (1991) An investigation into the relationship between gender and language, *Sex Roles* 24, 149–160.

Strauss, U., Grzybek, P., Altmann, **G.** (2007) Word length and word frequency. In: Grzybek, Peter (ed.): *Contributions to the science of text and language*. Dordrecht: Springer, 277–294.

Tannen, D. (1990) Gender differences in topical coherence: Creating involvement in best friends' talk. *Discourse Processes*, 13, 73–90.

Zipf, G. K. (1932) *Selected studies of the principle of relative frequency in language*. Cambridge, MA: Harvard University Press.

Zipf, G. K. (1935) *The psycho-biology of language: An introduction to dynamic philology*. Boston, MA: Houghton Miffin.

Appendix. Texts, authors and articles

Text	Title/Topic	Author
M01	American elections	Shane Goldmacher
F01	Britain's post-imperial fantasies	Jenny Clegg
M02	War in the Middle East	Tony Blair
F02	Prime minister's speech	Theresa May
M03	Computer assisted teaching/internet everywhere	Omar Mubin
F03	Student loans/children using internet	Jenny Adams
M04	Importance of humanities	Dan Hicks
F04	Transition to modernity	Rachel Seginer
M05	Teachers' versus parents' responsibility	Todd DeMitchell
F05	Teachers' versus parents' responsibility	Peggy Albers
M06	Stop bullying at school	Jonathan Todres
F06	Stop bullying at school	Emily Suski
M07	New teaching college	Sam Carr
F07	New teaching college	Samantha Twiselton
M08	Teacher crisis	Christopher Wilkins
F08	Teacher shortage	Kate Reynolds
M09	Climate change	Stéphane Boyer
F09	Climate change	Tara Martin
M10	Palestine-Israel peace process	Asaf Siniver
F10	Palestine-Israel peace process	Lori Allen
M11	Refugee crisis Greece	Dimitris Dalakoglou
F11	Refugee crisis Greece	Vicki Squire
M12	Refugee crisis Turkey	Durukan Kuzu
F12	Refugee crisis Turkey	Marianna Fotaki
M13	Elections compared	Scott Taylor
F13	Elections compared	Pippa Norris
M14	Presidential Election USA	Jesse Rhodes
F14	Presidential Election USA	Fiona Fidler
M15	Paintings forged in China	Martin Kemp
F15	Maya codices	Elizabeth Graham
M16	War art	Dan Peterson
F16	IVF holiday	Amy Speier
M17	Journalistic skills	Jimmy Smallwood
F17	Digital news	Mohadesa Najumi
M18	Journalistic skills	Jimmy Smallwood
F18	Unpaid internships	Danielle Cuaycong

Note: {Mxx, Fxx} denote matched male-female author pairs writing on the same topic.

Text	url
M01	https://www.gov.uk/government/speeches/britain-the-great-meritocracy-prime-ministers-speech
F01	https://theconversation.com/robots-likely-to-be-used-in-classrooms-as-learning-tools-not-teachers-66681
M02	https://theconversation.com/the-history-of-student-loans-goes-back-to-the-middle-ages-56326
F02	https://theconversation.com/heres-why-you-should-care-about-the-scrapping-of-a-level-anthropology-67332
M03	http://www.sciencedirect.com/science/article/pii/S0883035515000269
F03	https://theconversation.com/when-a-parent-directs-a-child-not-be-resuscitated-what-should-educators-do-55556
M04	https://theconversation.com/reading-to-your-child-the-difference-it-makes-57473
F04	https://theconversation.com/profiles/jonathan-todres-197891
M05	https://theconversation.com/profiles/emily-suski-241942
F05	https://theconversation.com/are-teachers-suffering-from-a-crisis-of-motivation-48637
M06	https://theconversation.com/college-of-teaching-will-be-an-opportunity-for-teachers-not-a-threat-to-their-independence-36237
F06	https://theconversation.com/hard-evidence-is-a-teacher-shortage-looming-34990
M07	https://theconversation.com/when-tackling-mediocre-schools-becomes-a-teacher-shortage-37384
F07	https://theconversation.com/climate-change-threatens-entire-ecosystems-lets-pick-them-up-and-move-them-57121
M08	https://theconversation.com/the-best-way-to-protect-us-from-climate-change-save-our-ecosystems-54110
F08	https://theconversation.com/israel-palestine-and-the-us-are-giving-up-on-the-peace-process-48458
M09	https://theconversation.com/us-is-the-real-obstacle-to-peace-between-israel-and-palestine-14139
F09	http://theconversation.com/raids-on-migrant-squats-in-greece-push-solidarity-efforts-further-to-the-margins-63421
M10	http://theconversation.com/welcome-to-city-plaza-athens-a-new-approach-to-housing-refugees-63904
F10	http://theconversation.com/turkey-is-buying-its-way-into-the-eu-with-a-deal-that-wont-solve-the-refugee-crisis-49331
M11	http://theconversation.com/outsourcing-a-humanitarian-crisis-to-turkey-is-that-the-european-thing-to-do-55915
F11	https://theconversation.com/can-quotas-make-gender-equality-happen-in-politics-lessons-from-business-65971
M12	https://theconversation.com/american-elections-ranked-worst-among-western-democracies-heres-why-56485
F12	https://theconversation.com/violence-has-long-been-a-feature-of-american-elections-67688
M13	https://theconversation.com/what-effect-will-closet-trump-voters-have-on-the-us-election-67928
F13	https://theconversation.com/you-may-spot-the-fake-at-dulwich-picture-gallery-but-forgeries-are-no-joke-24509
M14	https://theconversation.com/grolier-codex-ruled-genuine-what-the-oldest-manuscript-to-survive-spanish-conquest-reveals-67941
F14	http://theconversation.com/paul-nash-painted-in-the-trenches-and-i-did-the-same-in-afghanistan-67206
M15	https://theconversation.com/a-look-inside-the-czech-republics-booming-fertility-holiday-industry-52425
F15	http://www.huffingtonpost.co.uk/jimmy-smallwood/shorthand-journalism_b_12771762.html

M16	http://www.huffingtonpost.co.uk/mohadesa-najumi/the-new-age-of-digital-ne_b_12586592.html
F16	http://www.huffingtonpost.co.uk/jimmy-smallwood/shorthand-journalism_b_12771762.html
M17	http://www.huffingtonpost.co.uk/danielle-cuaycong-/unpaid-internships_b_12769708.html
F17	https://www.gov.uk/government/speeches/britain-the-great-meritocracy-prime-ministers-speech
M18	https://theconversation.com/robots-likely-to-be-used-in-classrooms-as-learning-tools-not-teachers-66681
F18	https://theconversation.com/the-history-of-student-loans-goes-back-to-the-middle-ages-56326