"He Said She Said" – a Male/Female Corpus of Polish

Filip Graliński*, Łukasz Borchmann†, Piotr Wierzchoń†

Adam Mickiewicz University
*Faculty of Mathematics and Computer Science / †Institute of Linguistics
*Umultowska 87, 61-614 Poznań, Poland / †al. Niepodległości 4, 61-874 Poznań, Poland
filipg@amu.edu.pl, borchmann@rainfox.org, wierzch@amu.edu.pl

Abstract

Gender differences in language use have long been of interest in linguistics. The task of automatic gender attribution has been considered in computational linguistics as well. Most research of this type is done using (usually English) texts with authorship metadata. In this paper, we propose a new method of male/female corpus creation based on gender-specific first-person expressions. The method was applied on CommonCrawl Web corpus for Polish (language, in which gender-revealing first-person expressions are particularly frequent) to yield a large (780M words) and varied collection of men's and women's texts. The whole procedure for building the corpus and filtering out unwanted texts is described in the present paper. The quality check was done on a random sample of the corpus to make sure that the majority (84%) of texts are correctly attributed, natural texts. Some preliminary (socio)linguistic insights (websites and words frequently occurring in male/female fragments) are given as well.

Keywords: Corpus Creation, Text Categorisation, Gender Attribution

1. Introduction

Differences in men's and women's language have been of interest in sociolinguistics for several decades (Lakoff, 1973), (Coates, 2004). The initial, sometimes hasty, conclusions drawn from relatively small collections of texts have been superseded by more reliable generalisations grounded in statistical analysis of large text corpora. For example, (Newman et al., 2008) analyse 14,000 English text samples (45M words) to find gender differences in language use, whereas (Lijffijt et al., 2014) worked with a subcorpus of the British National Corpus (409 texts, 16M words). The problem of gender bias in topics was studied and addressed in (Sarawgi et al., 2011), where corpora of blogs and scientific papers were used.

Large gender-tagged corpora for languages other than English are found rather sporadically; for instance, (Manjavacas, 2015) identifies differences in the mode of writing of Dutch men and women bloggers (831K posts, 114M words).

The corpora used in research on gender-related differences have been based on *metadata* supplied manually, for example, statements of gender by text authors, or gender tags added manually by the corpus creators. We propose a different approach: to look in the text itself for *gender-specific first-person expressions*. Not all languages have these (they are almost entirely absent in English), but they are quite frequent in the Slavic languages and they also occur, at least from time to time, in the Romance languages. For instance, *I am tired* will be rendered in Spanish as *Estoy cansado* or *Estoy cansada* depending on whether it was spoken or written by a man or a woman. Consequently, if a Spanish text fragment contains *Estoy cansado* or *Estoy cansada*, it might be assumed (without any metadata!) that the text was written by, respectively, a man or a woman, provided that we manage to exclude special cases such as quotations, formulaic expressions and metalinguistic discourse.

Our procedure is to take a very large Web corpus and grep for lines containing gender-specific first-person expressions to create a gender-specified subcorpus. We applied this procedure to Polish, a language in which the frequency of gender-specific first-person expressions is particularly high. In this paper, we describe the "He Said She Said" (HSSS) corpus obtained in this fashion.

2. Gender-specific First-person Expressions

A gender-specific first-person expression is a first-person expression which is specific to men or women due to grammatical constraints or requirements (usually related to the grammatical gender). It is usually a combination of words (e.g. *be* in the first person singular + masculine/feminine adjective), but in the case of Polish it might be a *single* word, as past verb forms are conjugated by gender; for instance, *I bought* is rendered as *kupiłem* by male speakers and *kupiłam* by female speakers.

See Table 1 for a comparison of gender-specific first-

person expressions across languages. We are not aware of any language in which the frequency of gender-specific first-person expressions is higher than in Polish. See Table 2 for examples of Polish gender-specific expressions.

3. Corpus Preparation

The starting point was the very large, freely available[1] Common Crawl-based Web corpus of Polish (Buck et al., 2014). In fact this is more of a Web "dump" than a coherent corpus. It contains 44.8G words (345 GB) extracted from 87M Web pages. The corpus is supplied as plain text, where each line is a text fragment (usually a single paragraph or sentence; total number of fragments is 9.9G). Meta-data (URL addresses) are given as separate lines.
To distil a He Said She Said subcorpus from the Common Crawl corpus, the following steps were taken:

1. Gender-specific inflected forms were extracted from PoliMorf (a morphological dictionary for Polish (Woliński et al., 2012)). We looked for either words constituting gender-specific first-person expressions on their own (like first-person past forms) or forms to be combined with specific words to make a gender-specific first-person expression. In the latter case, a regular expression was prepared manually to describe words which need to co-occur with a given inflected form (e.g. for masculine adjectives the regexp `b[eę]d[eę]|by[lł]em|by[lł]bym|jestem` was constructed to match the first-person singular forms of the verb *być =be*).

2. Ambiguous inflected forms were discarded (along with their masculine/feminine equivalents even if these were not ambiguous themselves), e.g. *podziałem* (past form of *podziać* or instrumental of the noun *podział*) was discarded along with its feminine counterpart (*podziałam*). In the case of some marginal ambiguities (i.e. when another interpretation was unlikely) the word was not discarded; for example *miałem* (the past form of *mieć=to have*) was kept, even though it might theoretically be a form of the noun *miał* (*coal dust*). The list of such exceptions was prepared manually.

3. Some special gender-specific first-person words were supplied manually (e.g. the irregular verb *powinienem/powinnam=I should*) and added to the list extracted from PoliMorf.

4. Gender-specific first-person expressions were found (by simple matching, without either lemmatisation or POS-tagging) in the Common Crawl corpus and marked. Text fragments without such expressions were discarded. For feminine forms, their masculine counterparts were also included as part of the tags (so that it would be easy to generate a gender-neutral version of the corpus – for instance, to create test sets for classifiers).

5. Text fragments containing both masculine and feminine first-person expressions were considered anomalies and were discarded (they were also logged for later inspection, as they would usually indicate an issue to be fixed in the procedure as a whole).

6. A significant number of text fragments consisted of website comments or message board posts (by more than one person) joined together (this was discovered following manual analysis of the logs described in the previous item). A regexp (matching time stamps and greetings) was handcrafted to isolate posts and only posts containing gender-specific first-person expressions were kept.

7. Text fragments matching a hand-crafted set of regular expressions were discarded. This step was intended to remove:

   - text from spammy sites (e.g. the expression *negacja logiczna=logical negation* occurred on many spam websites, probably due to an error in the spamming script);
   - formulaic expressions (e.g. "I have forgotten my password") – a list of such expressions was created by inspecting the most frequent fragments.

8. To avoid repetitions in the corpus only unique text fragments were kept (uniqueness was checked following aggressive normalisation involving lower-casing and removal of all non-letter characters and diacritics).

The entire procedure was repeated and refined a number of times. Each time a manual inspection of the results was performed and new filters and heuristics were implemented. For instance, during one round of inspection it was found that men were much more likely to use words such *hotel* or *room*; the real reason was that a large number of text fragments originated

[1] `http://data.statmt.org/ngrams/raw/`

	English	German	French	Polish	Russian
I + *be* + gender-specific noun	rare	✓	✓	✓	✓
I + *be* + masc./fem. adjective	×	×	✓	✓	✓
1st person past form	×	×	only some	✓(synthetically)	✓(analytically)
1st person conditional	×	×	×	✓(synthetically)	✓(analytically)
1st person future form	×	×	×	✓(optionally)	×

Table 1: Gender-specific first-person expressions in some European languages

type	English translation	Polish (male)	Polish (female)
I + *be* + gender-specific noun	*I am an actor/actress*	*jestem aktorem*	*jestem aktorką*
I + *be* + masc./fem. adjective	*I am tired*	*jestem zmęczony*	*jestem zmęczona*
1st person past form	*I sneezed*	*kichnąłem*	*kichnęłam*
1st person conditional	*I would sneeze*	*kichnąłbym*	*kichnęłabym*
1st person future form	*I will be dancing*	*będę tańczył*	*będę tańczyła*

Table 2: Examples of Polish gender-specific first-person expressions

from the Polish version of the TripAdvisor website[2] in which machine translation was used to translate foreign (mostly English) reviews into Polish. Masculine first-person forms were always used in these automatic translations, even if the reviewer was female. Following, the decision was made simply to remove TripAdvisor text fragments from the HSSS corpus.

The tools and scripts used for creating the HSSS corpus out of the Common Crawl corpus are available at `git://gonito.net/petite-difference`. The corpus itself is available at `http://mrt.wmi.amu.edu.pl:8888/gender-subcorpus-pl.txt.gz`.

3.1. Asymmetric gender-specific forms

A number of *asymmetric* gender-specific forms were found, i.e. expressions which indicate a female speaker when the feminine form is used, but whose masculine form could be used by either a male or a female. For instance, past first-person plural masculine forms are used for both masculine and mixed gender groups (whereas the feminine form is used only for all-female groups).

We propose the following treatment of asymmetric expressions when creating the HSSS corpus:

- obviously, their masculine forms cannot be used for marking texts as authored by a male (as they are used by females in some contexts as well),
- what is less obvious is that their feminine forms should not be used either (otherwise, an artificial imbalance in such forms would be introduced, as they would be more likely to occur in the feminine part of the final corpus),
- on the other hand, feminine forms should be marked as such and normalised into masculine forms when, for instance, gender classification challenge is prepared.

Some asymmetric forms are due to accidental homonymy, for instance, *piekłam* is unambiguously the feminine form of the verb *piec* (= *bake*), whereas *piekłem* is either the masculine form of this verb or an inflected form of the noun *piekło* (= *hell*). Such expressions should be handled just in the same way as other asymmetric forms.

4. Corpus Statistics

Basic statistics for the corpus are given in Table 3.

A random sample of 1,261 text fragments was manually checked for the following types of anomalies:

1. The marking of gender-specific words is incorrect (not all gender-specific words are marked, or some words are incorrectly marked).
2. The fragment contains a significant part written or spoken by a person whose gender is unknown or differs from that of the person who used the gender-indicating first-person forms (for example, an alternation of comments by two people on an Internet forum, where the gender of only one of them can be determined on the basis of first-person forms).

[2] `http://pl.tripadvisor.com/`

	bytes	words	fragments
total	5,421,911,918	781,772,768	15,866,774
male	3,426,844,921 (63.2%)	492,144,270 (63.0%)	9,910,946 (62.5%)
female	1,995,066,997 (36.8%)	289,628,498 (37.0%)	5,955,828 (37.5%)

Table 3: Corpus statistics

3. The first-person form appears in a quotation (reported by a journalist, for instance) or in the speech or thoughts of a character in a literary text (memoirs and autobiographies are not subject to this problem, except where they include the utterances of other persons besides the author). In particular, erotic or pornographic fiction was marked in this category.

4. The first-person form is part of a title (e.g. the title of the Polish film *Jak rozpętałem II wojnę światową*).

5. The gender-specific word is part of a formula typically used on websites and forms (e.g *Zapomniałem hasła=I have forgotten the password*).

6. The text is of artificial origin, is not a natural Polish text (random sequence of words generated by a spammer, the result of machine translation, lists of words, a collection of headers).

7. Most of the fragment is in a foreign language (this problem does not concern short phrases or titles expressed in a foreign language).

The distribution of anomalies is given in Table 4.

	number of occurrences	percentage
none	1064	84.3%
(1)	14	1.1%
(2)	123	9.7%
(3)	41	3.2%
(4)	0	0.0%
(5)	0	0.0%
(6)	15	1.1%
(7)	5	0.3%

Table 4: Distribution of anomalies in the random sample

5. Insights and Applications

The HSSS corpus can give interesting insights into gender-related differences in language and social life. For instance, see Table 5 and Table 6 for the lists of words for which the largest imbalance between men's and women's usage was found. Another example is given in Table 7. Of course, one needs to be very careful when interpreting data extracted from the HSSS corpus (is this just an artefact of the procedure? is this due to some systematic noise?). For instance, the word *zadowolona* in Table 6 turns out to a "leaked" feminine first-person form: it occurred frequently as part of the expression *jestem bardzo zadowolona* (*I am very glad*), which was not covered by regular expressions described in Section 3. (as the adjective *zadowolona* was not adjacent to the form of verb *be*).[3]

The HSSS corpus was not balanced between topics (as can be seen in Table 7), which might be desirable for some sociolinguistic and extra-linguistic studies, but which usually poses a problem for purely linguistic research. Let us note, however, that for each text, its source (URL address) is given in the HSSS corpus and a subcorpus balanced between websites (which be treated as an approximation of topics) could be created.

Another application of the HSSS corpus is to train and test classifiers for guessing the gender of a text author. We intend to prepare and release such data sets based on the HSSS corpus and balanced against websites.

6. Future Work

The method described in this paper could be applied for other languages, e.g. for French, in which, on one hand, the frequency of gender-specific first-person forms is much lower than in Polish, but which, on the other hand, has twice as large CommonCrawl corpus.

We plan to use gender-specific first-person forms for the diachronic study of Polish language. For instance, see Figure 1 for the graph of the relative frequency of feminine/masculine first-person forms obtained on a large, time-extensive corpus of Polish (Graliński, 2013) using the scripts initially created for the HSSS corpus.

[3]This particular "leak" will be fixed in the next release of the corpus

	word		$f_M/(f_F+2000)$	f_M	f_F
1.	silnik	engine	14.85	56974	1837
2.	windows	Windows	12.33	61660	2999
3.	silnika	engine (gen.)	11.23	35907	1197
4.	gb	GB	10.64	42441	1987
5.	mb	MB	9.57	38266	2000
6.	meczu	(football) match	9.44	55055	3835
7.	pc	PC	8.70	31735	1649
8.	opony	tires	8.54	35132	2116
9.	apple	Apple	8.50	29362	1454
10.	iphone	iPhone	8.25	26022	1156
11.	zwiastuny	(movie) trailer	7.94	20249	549
12.	hd	HD	7.92	26082	1295
13.	ubuntu	Ubuntu	7.86	19573	489
14.	systemu	system (gen.)	7.85	62447	5953
15.	serwer	server	7.76	27968	1606

Table 5: The words with the highest male/female imbalance.

	word		$f_F/(f_M+2000)$	f_F	f_M
1.	ciąży	pregnancy (gen.)	13.08	105203	6042
2.	miesiączki	menstruation (gen.)	11.03	25966	354
3.	ciasto	dough	10.44	73753	5065
4.	ciążę	pregnancy (acc.)	8.68	29532	1401
5.	zadowolona	glad	8.34	40416	2846
6.	ciąża	pregnancy	7.91	24008	1036
7.	ciazy	pregnancy (gen.)	7.70	21483	791
8.	antykoncepcyjne	contraceptive	7.67	18346	393
9.	ginekologa	gynaecologist (gen.)	7.66	19497	544
10.	tabletki	pills	7.26	68851	7490
11.	porodzie	childbirth	6.63	18662	813
12.	mąż	husband	6.30	56927	7029
13.	miesiączkę	menstruation (acc.)	6.16	12641	53
14.	krwawienie	bleeding	5.96	16921	839
15.	ciasta	dough (gen.)	5.94	30994	3214

Table 6: The words with the highest female/male imbalance.

Acknowledgements

Work supported by the **Polish Ministry of Science and Higher Education** under the **National Programme for Development of the Humanities**, grant 0286/NPRH4/H1a/83/2015: "50 000 słów. Indeks tematyczno-chronologizacyjny 1918-1939".

7. References

Buck, Christian, Kenneth Heafield, and Bas van Ooyen. 2014. N-gram counts and language models from the common crawl. In *Proceedings of the Language Resources and Evaluation Conference*, Reykjavk, Icelandik, Iceland, May.

Coates, Jennifer. 2004. *Women, men, and language: A sociolinguistic account of gender differences in language*. Pearson Education.

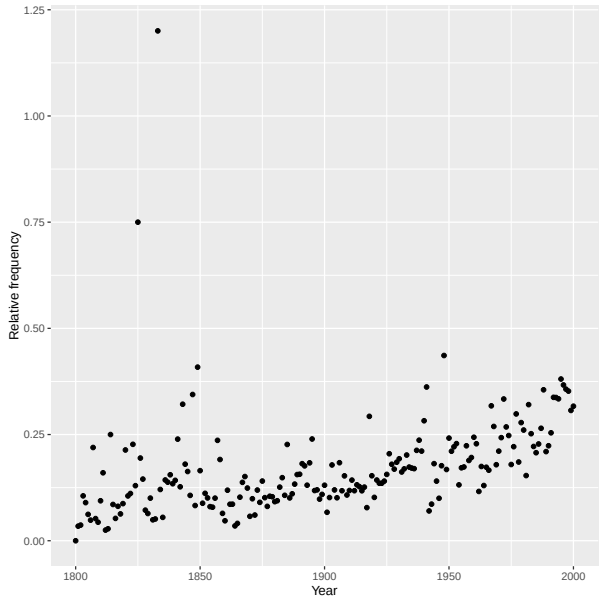
Graliński, Filip. 2013. Polish digital libraries as a text corpus. In Vetulani, Zygmunt and Hans Uszkoreit, editors, *Proceedings of 6th Language & Technology Conference*, Poznań. Fundacja Uniwersytetu im. Adama Mickiewicza.

Lakoff, Robin. 1973. Language and woman's place. *Language in society*, 2(01):45–79.

	website	#M	website	#F
1.	www.youtube.com	323318	www.youtube.com	103234
2.	www.sfd.pl	149098	forum.gazeta.pl	65881
3.	www.wykop.pl	124325	gwiazdunie.pl	56315
4.	www.kfd.pl	87873	www.photoblog.pl	47771
5.	www.elektroda.pl	84414	szafa.pl	43758
6.	www.autocentrum.pl	67767	www.kotek.pl	38742
7.	www.dobreprogramy.pl	59990	parenting.pl	35037
8.	flaker.pl	47259	www.forum-turystyczne.pl	32572
9.	myapple.pl	46135	www.babyboom.pl	29489
10.	forum.gazeta.pl	44633	tematy.abcciaza.pl	29267

Table 7: The most popular websites among men and women

Figure 1: Relative frequency of feminine/masculine first-person forms



Lijffijt, Jefrey, Terttu Nevalainen, Tanja Saily, Panagiotis Papapetrou, Kai Puolamaki, and Heikki Mannila. 2014. Significance testing of word frequencies in corpora. *Digital Scholarship in the Humanities*.

Manjavacas, Enrique. 2015. Statistical description of gender-related differences in language use in a dutch blog corpus.

Newman, Matthew L, Carla J Groom, Lori D Handelman, and James W Pennebaker. 2008. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45(3):211–236.

Sarawgi, Ruchita, Kailash Gajulapalli, and Yejin Choi. 2011. Gender attribution: Tracing stylometric evidence beyond topic and genre. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, CoNLL '11, Stroudsburg, PA, USA. Association for Computational Linguistics.

Woliński, Marcin, Marcin Miłkowski, Maciej Ogrodniczuk, Adam Przepiórkowski, and Łukasz Szałkiewicz. 2012. PoliMorf: a (not so) new open morphological dictionary for Polish. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*, Istanbul, Turkey. ELRA.