УДК 811.111-26

Василевич А.П., Мамаев М.М.
Московский государственный областной университет

ПРОБЛЕМА ВЫДЕЛЕНИЯ ГЕНДЕРНО ЗНАЧИМЫХ ПАРАМЕТРОВ ТЕКСТА

Аннотация. Гендерные особенности языка проявляются прежде всего на лексическом уровне. При этом основным исследовательским приёмом является оценка употребительности: одни слова чаще встречаются в текстах, написанных женщинами, другие – в текстах мужчин. Основанная на этом подходе процедура гендерной атрибуции текста была опробована на обширном англоязычном материале (22 автора) и оказалась достаточно эффективной. Расположение авторов на «шкале маскулинности» позволяет ставить и решать ряд нетривиальных задач. В частности, выдвинута гипотеза о том, что по сравнению с авторами XIX века степень маскулинности современных авторов-мужчин в целом значительно снизилась, а авторов-женщин – повысилась.
Ключевые слова: английский художественный текст, гендерные особенности, гендерная атрибуция текста, употребительность слова.

A. Vasilevich, M. Mamaev
Moscow State Regional University

WAY TO DETECT GENDER INHERENT TEXT ELEMENTS

Abstract. Gender linguistic characteristics become apparent essentially at the lexical level. So evaluating frequency of use becomes the main research method: one set of words is more frequent in the texts written by women, others are more frequent in men's texts. The procedure of gender attribution based on this approach was tested on a large sample of English texts (22 authors) and proved to be quite effective. The placement of authors on the "scale of masculinity" allows us to put forward and solve a number of nontrivial problems. In particular, the hypothesis was advanced that compared to the 19th century authors the level of masculinity of contemporary male authors has decreased while the level of masculinity of female authors increased.
Key words: English fiction, gender peculiarities, gender text attribution, word frequency evaluation.

В настоящее время во многих науках наметилась тенденция к более полному изучению человека, причём интерес представляет не просто человек, а прежде всего конкретный человек как носитель сознания и языка. В этой связи, в частности, обрели большое значение вопросы пола, которые представляют интерес не только в научном плане, но в социокультурном отношении. При этом важным является разделение понятий «пол» и «гендер».

«Пол» рассматривается как биологическое явление; соответственно, мужчины противопоставляются женщинам по чисто биологическим признакам, включая поведенческие осо-

бенности. «Гендер» же затрагивает в первую очередь свойства психики, а также социокультурные образования, складывающиеся в ходе общественного развития.

Гендерная идентичность не даётся индивиду автоматически, при рождении, а вырабатывается в результате сложного взаимодействия его природных задатков и соответствующей социализации. Через игры, одежду и т. п. ребёнок с малолетства начинает себя идентифицировать либо с мужским, либо с женским началом. Изучение характерных особенностей мужчин и женщин является предметом новой науки – гендерологии. Комплекс мужских черт получил название «маскулинность», комплекс женских – «фемининность» («женственность») [2, с. 75–80].

Естественно предположить, что поскольку гендер – это вполне определённый комплекс сложившихся социальных и психологических установок, он воздействует на *языковое поведение* личности [4, с. 124].

Цель данной работы – рассмотреть связь гендера с языковой личностью.

Введение категории «гендер» в исследовательский аппарат лингвистики открыло новые перспективы для анализа различных аспектов языка и речи. Сам термин появился в лингвистике в 80-е годы прошлого века[1].

К настоящему времени сформировались несколько лингвистических направлений, различающихся по концептуальным установкам, методам исследования и характеру изучаемого материала [1]. Пусть и под разным углом зрения, эти направления изучают, в сущности, всего две большие группы проблем:

1. Собственно язык и отражение в нём гендерного фактора: номинативная система, лексикон, синтаксис, категорию рода и ряд сходных объектов. Эти исследования могут быть нацелены на какой-то один конкретный язык, а могут привлекать и сопоставительный анализ.

2. Речевое поведение мужчин и женщин, с выделением типичных стратегий и тактик, специфического гендерного выбора единиц лексикона и т. д. – то есть выявление специфических черт мужской и женской речи.

Ряд конкретных гендерных языковых особенностей описала Р. Лакофф [7]. Она отмечает, например, что прилагательные цвета принадлежат к активному словарю женщин и практически отсутствуют в вокабуляре мужчин; женщинам свойственны аффективные прилагательные и употребление частиц, выражающих эмоции (*Oh! Ah!*) и т. д.

В этой и других подобных работах языковые особенности выделялись и на синтаксическом, и на морфологическом уровне, но наиболее показательными они являются **на лексическом уровне**. Соответственно, основным исследовательским приёмом является оценка употребительности: слова или лексические обороты объявляются «фемининными», если они существенно чаще встречаются в произведениях авторов-женщин. Следовательно, самое пристальное внимание здесь требуется уделять *частотности* языковых единиц.

Интересно было бы рассмотреть об-

[1] Основополагающими публикациями здесь можно считать монографии Робин Лакофф [7] и Дженнифер Коутс [6]. Краткий обзор работ см. в: [3].

ратную задачу: можно ли в результате лексического анализа текста с той или иной степенью успеха осуществить гендерную атрибуцию, т. е. определить, был ли автор текста мужчиной или женщиной. Например, установлено, что мужчинам свойственна повышенная частотность жаргонизмов или латинских терминов, но ведь жаргонизмы и латинские термины присутствуют далеко не в *каждом тексте*. Равно как не в каждом феминном тексте можно встретить эвфемизмы и разделительные вопросы. Хотелось бы, чтобы в любом наугад выбранном тексте присутствовали такие элементы, которые позволили бы достаточно надёжно провести гендерную атрибуцию. Напрашивающимся решением здесь представляется обращение к **служебным словам**.

Прежде всего, служебные слова по самой своей природе очень употребительны и встречаются *в любых текстах*. Не маловажно и то обстоятельство, что их относительно немного, и, значит, сам процесс анализа не будет слишком трудоёмким.

Следует заметить, что гендерный характер служебных слов практически не исследован. Нам известна лишь одна попытка решить сходную задачу на английском материале[1]. Правда, речь в работе идёт о классе функциональных слов (*function*, или *functional words*), который включает не только служебные слова, но ещё местоимения и модальные глаголы (всего 30 слов). Зато авторы пытаются решить именно задачу гендерной атрибуции: сначала делят функциональные слова на преимущественно «мужские» и преимущественно «женские», а потом предлагают алгоритм, который на основе подсчёта употребительности выбранной группы слов позволяет с высокой вероятностью установить пол автора текста.

Позднее предложенный исследователями алгоритм расчёта был испытан на обширном корпусе текстов с чёткой гендерной принадлежностью [5]. Было показано, что алгоритм позволяет определить гендерную атрибуцию автора текста с восьмидесятипроцентной вероятностью. Этот процент представляется нам достаточно высоким, причём очевидно, что остаётся возможность дальнейшего совершенствования процедуры применения коэффициента.

Согласно алгоритму, артикли *a, the*, а также указательные слова *that, these* являются маскулинными показателями, в то время как группа местоимений *I, you, she, her, their, myself, yourself, herself* указывают на принадлежность текста автору-женщине. В основе действия алгоритма лежит прежде всего частота каждого функционального слова из заданного списка. Данные о частоте дополняются системой коэффициентов, которая регулирует «вклад» каждого слова в конечный результат. Например, коэффициент, или «вес» предлога *with* – 52, «вес» местоимения *who* – 19, артикля *the* – 7 и т. д. Таким образом, если *with* встретилось в тексте 4 раза, то его суммарный «вклад» составит (4 × 52) – 208; если *the* встретилось 69 раз, то его вклад будет весьма весомым (69 × 7) – 483.

Отдельно подсчитывается сумма «весов» для «феминных» слов (*with,*

[1] См.: Koppel M., Argamon S., Shimoni A.R. Automatically determining the gender of a text's author // Bar-Ilan University Technical Report BIU-TR-01-32. – 2001.

if, not и т. д.) и «маскулинных» (*who, the, as* и т. д.), и полученные суммы сравниваются. Если итоговая сумма феминных слов оказывается больше итоговой суммы группы маскулинных слов, то текст атрибутируется как «фемининный».

Хотя по утверждению исследователей алгоритм функционирует достаточно эффективно, предложенный ими набор функциональных слов вызывает у нас определённые сомнения. Так, непонятно, почему авторы включают в список формы глаголов *to be* и *to say*. Вызывает вопросы и использование *to*: ясно, что очень большая разница, выступает ли *to* в качестве частицы с инфинитивом, или в качестве предлога (при подсчётах эта функциональная разница никак не учитывается). Исходя из сказанного, мы решили проверить действие алгоритма на новой выборке текстов.

Для анализа было отобрано 11 произведений британских авторов XIX – начала XX вв. (J. Austen, Ch. и A. Brontë, Ch. Dickens, G. Chesterton и др.) и 11 романов XX-XXI вв. (J.K. Rowling, J. Cox, K. Brooks, P. Ness и др.). Каждая эпоха была представлена примерно равным количеством мужчин и женщин-авторов. Принимая во внимание характер материала (функциональные слова), было излишне анализировать тексты в полном объёме. Мы ограничились анализом трёх отрывков из каждого произведения – соответственно из начала, середины и конца. Объём каждого отрывка во всех случаях был примерно одинаковым – порядка 1500 слов.

Обработка текстов производилась с помощью алгоритма *Gender Genie*, размещённого в сети Интернет (онлайн-сервис *Gender Genie* [8]). Алгоритм позволяет любому пользователю интернета предъявить свой текст (определённого объёма), и программа автоматически обработает его, выделив слова, релевантные для гендерной принадлежности авторов.

Пример исходной матрицы результатов представлен в табл. 1.

Таблица 1

Исходная матрица результатов подсчётов (фрагмент)

	«Феминные» слова				«Маскулинные» слова			
	Слово	Частота	«Вес» по алгоритму	Общий вклад	Слово	Частота	«Вес» по алгоритму	Общий вклад
Charles Dickens	her	27	20	540	the	209	6	1254
	me	26	20	520	as	43	30	1240
	…	…	…	…	…	…	…	…
	with	43	1	43	many	4	6	24
	where	3	2	6	more	7	2	14
				Всего	**4179**			Всего	**5456**
Ann Brontë	her	61	20	1220	the	170	6	1020
	not	65	8	520	as	32	30	960
	…	…	…	…	…	…	…	…
	with	51	1	51	these	1	8	8
	where	3	2	6	many	1	6	6
				Всего	**5400**			Всего	**4939**

Согласно данным табл. 1 в тексте Диккенса преобладает сумма маскулинных слов (5456 > 4179); стало быть, этот текст написан автором-мужчиной. Напротив, в тексте Бронтё преобладают феминные слова (5400 > 4939), значит, текст написан женщиной. В обоих случаях программа не ошиблась.

Однако, когда мы проанализировали тексты всех 22 авторов, результат оказался не столь впечатляющим. Из 11 текстов, написанных авторами-мужчинами, программа верно атрибутировала всего 6, т. е. ошиблась почти в половине случаев. Если не отвергать саму идею о возможности гендерной атрибуции текста на основе ограниченного числа слов, то приходится признать, что либо неудачными являются некоторые слова из рассматриваемого списка, либо требуется корректировка системы «вéсов», либо верно и то, и другое.

Наш следующий шаг имел две основных цели:

1) убедиться в правильности изначального деления слов на «маскулинные» и «феминнные»; при необходимости произвести изменения в составе соответствующих групп;

2) совершенствовать систему «вéсов» используемых слов.

Мы вернулись к исходным матрицам и обобщили данные по всем 22 авторам. Анализ показал, что в группе феминнных слов подтверждающие данные получены для 10 слов из 15 (например, *me* встретилось у женщин в общей сложности 360 раз, у мужчин – только 180 и т. д.). Для четырёх слов явных предпочтений не выявлено (ср.: *with* – 360 : 360, *when* – 160 : 150), а в одном случае вообще был получен обратный результат (*was* – 860 : 1033).

Ещё более обескураживающим оказался результат для маскулинных слов. Здесь подтверждение зафиксировано всего для 6 слов.

Естественно было устранить из рассмотрения те слова, которые не выявили чётких гендерных тенденций. В результате в списке осталось 10 «феминнных» и 7 «маскулинных» слов. Следующая наша задача – скорректировать систему «весов» отобранных 17 слов.

Логика приписывания «весов» авторами алгоритма была нам не понятна. Приведём несколько примеров (табл. 2).

В группах (1, 2, 3) собраны слова примерно одинаковой употребитель-

Таблица 2

Анализ «веса» слов, предложенных алгоритмом

№ п/п	Слово	Встретилось в текстах авторов-мужчин	Встретилось в текстах авторов-женщин	Вес слова
1	should	32	48	50
	around	43	49	10
	myself	37	51	4
2	we	102	169	45
	what	136	190	35
3	a	1347	1111	10
	to	1083	1240	2

ности. Как мы видим, «веса» слов в каждой группе различаются (что называется – в разы). Необходимость в корректировке системы «весов» была для нас очевидна.

При создании новой системы «весов» мы приняли во внимание два основных фактора. Первый из них – величина разницы в частоте. В табл. 2 слова группы 2, например, обладают заметно бóльшей дифференцирующей силой, чем слова группы 1, и им следовало бы соответственно придать бóльший вес.

С другой стороны, некоторые слова (особенно артикли) имеют очень высокую употребительность, и придание им большого веса приведёт к тому, что они возьмут на себя решающую долю общего показателя, сведя к минимуму роль остальных слов. Мы сочли целесообразным при определении веса добиться того, чтобы доля одного слова в общей сумме не превышала 20%. Так, артиклю *the* мы присвоили вес 4. В то же время *if* получило вес 28, а *myself* – 50.

Далее мы решили усовершенствовать процедуру ещё в одном отношении. Алгоритм, который мы взяли за основу, позволяет лишь квалифицировать текст как «мужской» или «женский». Нам хотелось бы пойти дальше, расположив тексты на своеобразной «шкале маскулинности». Это позволит не только отличить «мужской» текст от «женского», но и сопоставить между собой два мужских или два женских текста. В результате мы разработали статистический показатель маскулинности **М**, который может служить важной характеристикой гендерного стиля автора. Показатель **М** напрямую связан с долей «маскулинных» слов и вычисляется по формуле:

$$\mathbf{M}_{ai} = \frac{\text{Маск}_{ai}}{\text{Ср}_{маск}} - \frac{\text{Фем}_{ai}}{\text{Ср}_{фем}}$$

где $\mathbf{M}_{ai}$ – показатель степени маскулинности автора (i); Маск_{ai} – суммарная доля всех маскулинных слов автора (i); Фем_{ai} – суммарная доля всех феминных слов автора (i); Ср – доля маскулинных (и, соответственно, феминных) слов в среднем по всей группе текстов.

Приведём пример расчёта для текстов Честертона и А. Бронтё. В табл. 3 представлен фрагмент исходной матрицы. Применив к этим данным формулу расчёта показателя **М**, мы получили следующие результаты: $\text{М}_{\text{Честертон}} = + 0{,}809$; $\text{М}_{\text{Бронтё}} = - 0{,}763$.

Понятно, что чем выше показатель **М**, тем «маскулиннее» соответствующий автор. Отрицательные значения **М** предположительно означают, что автор – женщина. Насколько это предположение соответствует фактам – показывает табл. 4, в которой представлены результаты подсчётов по всем текстам.

Как видно из табл. 4, введённый нами коэффициент достаточно точно отражает реальную гендерную принадлежность автора. Конечно, как это всегда бывает в таких случаях, имеются пограничные случаи, когда **М** писателя близок к 0 (к этой категории можно отнести Диккенса, Теккерея и Брукса у мужчин, а также Роулинг, Кокс, Бронтё и Розофф – у женщин). Возможно, при дальнейшем совершенствовании процедуры удастся уменьшить число подобных примеров, однако следует заметить, что более перспективным нам представляется другое направление исследований.

Как мы говорили выше, гендер – это отражение языкового сознания, и он не во всех случаях обязан совпадать с

Таблица 3

Исходная матрица данных (фрагмент)

	Феминные слова				Маскулинные слова			
	Слово	Частота	Вес	Суммарная доля слова	Слово	Частота	Вес	Суммарная доля слова
Честертон	we	7	40	280	the	292	4	1168
	myself	4	50	200	as	44	20	880
	…			…	…			…
	she	1	12	12	around	0	20	0
	Всего			**1585**	Всего			**4572**
Бронтё	her	61	15	915	the	170	4	680
	myself	17	50	850	a	86	6	516
	…			…	…			…
	be	17	10	170	around	2	20	40
	Всего			**5515**	Всего			**2684**

Таблица 4

Показатель маскулинности M всей группы авторов

Авторы-мужчины	**M**	Авторы-женщины	**M**
O. Wilde	0,83	J.K. Rowling	0,09
G.Chesterton	0,81	Jennifer Cox	– 0,02
R. Stevenson	0,49	Meg Rosoff	– 0,03
Bill Bryson	0,39	Ch. Bronte	– 0,04
Conan Doyle	0,34	Mary Shelley	– 0,15
Patrick Ness	0,29	A. Christie	– 0,19
Mal Peet	0,25	J. Valentine	– 0,41
M. Burgess	0,19	Anne Fine	– 0,47
Ch. Dickens	– 0,02	M. Blackman	– 0,67
W. Thackeray	– 0,06	Jane Austen	– 0,74
Kevin Brooks	– 0,12	Ann Bronte	– 0,76

биологическим полом. Соответственно, если **M** данного автора выбивается из общего ряда, это должно порождать определённое сомнение в его гендерной принадлежности. Тогда потребуется привлечь дополнительные факты из смежных областей знаний – литературоведения, психологии и т. д. С этой точки зрения было бы интересным рассмотреть подробнее результат Кевина Брукса или же результат сестер Бронтё, которые оказались практически на противоположных концах шкалы маскулинности.

В заключение обратим внимание ещё на один интересный факт. Наша

выборка художественных текстов была сбалансирована не только по полу авторов (11 мужчин и 11 женщин), но и по времени создания (11 текстов XIX – начала XX вв. и 11 текстов XXI в.). Анализ показал, что у писателей XIX в. гендерная дифференциация была отчётливее ($М_{мужчин}$ в среднем = + 0,34; а $М_{женщин}$ = – 0,40, т. е. разность составила 0,74). У современных авторов намечается тенденция к сближению ($М_{мужчин}$ = + 0,25, $М_{женщин}$ = – 0,21; разность – всего 0,45). Причём этот процесс «двусторонний»: мужчины стали менее «маскулинными», а женщины – более «маскулинными».

Отметим, что факты такого рода, конечно, требуют дополнительной проверки на новом материале.

ЛИТЕРАТУРА:

1. Гвоздева А.А. Языковая картина мира: лингвокультурологические и гендерные особенности (на материале художественных произведений русскоязычных и англоязычных авторов): автореферат дис. … канд. филол. наук. – Краснодар, 2004. – 15 с.
2. Зыкова И.В. Актуальные проблемы изучения гендерного фактора в рамках лингвистических исследований (на материале английского языка) // Межкультурная коммуникация и перевод. Материалы межвузовской конференции. – М.: МОСУ, 2002. – 210 с.
3. Мамаев М.М. Многоаспектный характер изучения гендерного фактора в лингвистике // Вестник Московского государственного областного университета. Серия «Лингвистика». 2011. – № 2. – С. 32–37.
4. Маслова В.А. Лингвокультурология: Учеб. пособие для студ. высш. учеб. заведений. – 2-е изд., стереотип. – М.: Издательский центр «Академия», 2004. – 208 с.
5. Argamon S. et al. Gender, Genre, and Writing Style in Formal Written Texts / Argamon S., Koppel M., Fine J., Shimoni A.R. // Interdisciplinary Journal for the Study of Discourse. – 2006. – Vol. 23, Issue 3. – P. 321–346.
6. Coates J. Women, Men and Language: A Sociolinguistic Account of Gender Differences in Language. – Harlow England, New York: Longman, 2004. – 254 p.
7. Lakoff R. Language and Woman's Place. – N.Y.: Harper and Row, 1975. – 85 p.
8. The Gender Genie. [Электронный ресурс]. – URL: http://bookblog.net/gender/genie.php (дата обращения: 19.02.2013).