

УДК 81'33
ББК 81.1

Т.А. Литвинова

К ПРОБЛЕМЕ СТАБИЛЬНОСТИ ХАРАКТЕРИСТИК ИДИОСТИЛЯ*

Представлены результаты пилотного исследования стабильности/изменчивости количественных параметров текстов, представляющих собой образцы естественной русской письменной речи, с использованием русскоязычной версии программы *Linguistic Inquiry and Word Count*, и показана возможность использования этих результатов для решения автороведческих задач.

Ключевые слова: *стилеметрия, автороведение, идиостиль, письменная речь, атрибуция текста, корпус текстов, стабильность и вариативность параметров текста.*

Литвинова Татьяна Александровна – канд. филол. наук, научный сотрудник Регионального центра русского языка Воронежского государственного педагогического университета
Тел.: 8-980-342-00-73
E-mail: centr_rus_yaz@mail.ru

© Литвинова Т.А., 2015.

* Исследование выполнено при поддержке гранта РФФИ № 15-36-50104 «Формально-грамматические параметры письменного текста и индивидуально-личностные характеристики его автора: корпусное исследование», гранта РГНФ № 15-34-01221 «Детекция лжи в письменном тексте: корпусное исследование».

Введение. В многочисленных исследованиях по идентификации автора на основе анализа количественных параметров его письменных текстов было доказано существование у каждого пишущего (не только у профессиональных писателей) своеобразного «языкового отпечатка», формирующего его идиостиль. Этот «языковой отпечаток» получил название «стилом» (ср. геном) [New machine learning methods...]. Предполагается также, что люди, объединенные каким-либо общим признаком (принадлежностью к одному и тому полу, наличием сходных психологических характеристик и т.д.), имеют схожие стиломы, благодаря чему становится возможной разработка методик диагностирования по тексту тех или иных личностных характеристик (см., например, [Koppel, Argamon, Shimon, p. 401 – 412], а также краткий обзор в [Литвинова 2013а, с. 64 – 72]).

Авторы работ по идентификации и диагностированию личности автора текста исходят из предположения об устойчивости, стабильности анализируемых ими параметров текста. Вместе с тем известно, что определенные характеристики идиостиля меняются с возрастом [Pennebaker, Stone, p. 291 – 301; Андреев], в стрессовых ситуациях [Pennebaker, Lay, p. 271 – 272; Носенко] и т.д.

Таким образом, анализ научных работ по атрибуции текста и диагностированию как устойчивых характеристик, так и временных состояний автора текста позволяет предположить наличие в идиостиле каждого продуцента текста как стабильных, так и ва-

рырующихся в зависимости от различных причин языковых параметров. Несмотря на очевидную теоретическую и практическую значимость проблемы соотношения устойчивых и изменчивых элементов в идиостиле, как справедливо отмечает В.С. Андреев, этот вопрос остается малоизученным: до сих пор известно крайне мало работ, системно описывающих изменения идиостиля и того, какие именно его характеристики обладают значительной устойчивостью [Андреев]. Кроме того, имеющиеся отдельные работы выполнены на материале художественных произведений, проблема же устойчивости и изменчивости параметров идиостиля рядовых носителей языка практически не исследована.

Так, в работе В.П. и Т.Г. Фоменко [Фоменко, Фоменко, с. 688 – 722] была предпринята попытка выявления так называемого **авторского инварианта**, т. е. такого параметра текста, значения которого стабилизируются на больших отрезках текстов, но при этом позволяют различать тексты разных авторов. Оказалось, что среди 9 формализуемых параметров текста (длина предложений в словах; длина слов в слогах; частота употребления существительных, глаголов, прилагательных и т.д.) только параметр «доля служебных слов» подходит под эти требования. Все остальные параметры либо не стабилизируются, либо не различают авторов. Интересно, что обнаруженный учеными на обширном языковом материале (произведения 27 писателей и поэтов) авторский инвариант практически не зависел от временной эпохи, к которой принадлежал автор текстов, – исследования проводились на материале текстов писателей трех веков – от XVIII до XX.

В работе В.С. Андреева, посвященной анализу динамического аспекта идиостиля, выделяются следующие группы параметров текстов: 1) **константы высшего уровня** – маркеры стиля, которые не различают стили разных авторов в таксономическом плане и при этом не изменяются в процессе творчества у каждого автора во времени; 2) **авторский стилем** – группа признаков, которые различают стили авторов, но не меняются в творчестве каждого из них со временем, входя в интегральную модель индивидуального стиля данного автора (ср. «авторский инвариант»); 3) **признаки комонотонного варьирования** – кластер характеристик, которые не различают стили авторов, при этом значимо меняясь у каждого из них от одного этапа творчества к другому; 4) **зона индивидуального поиска** – признаки, значимо изменяющиеся во времени у каждого автора и дифференцирующие стили этих авторов [Андреев].

В нашей работе мы ставим задачу выявления стабильных и вариативных характеристик идиостиля рядовых носителей языка на материале их естественной письменной речи.

Материал и методы. В качестве материала исследования были использован специально созданный для решения автороведческих задач корпус текстов *Personality* [Загоровская, Литвинова, Литвинова, с. 387 – 389], содержащий, помимо собственно текстов, метаразметку в виде данных об их авторах (пол, возраст, образование, данных психологиче-

ского тестирования и др.). В настоящее время корпус насчитывает 1 867 текстов от 1 145 респондентов и постоянно пополняется.

Для анализа текстов нами была использована разработанная американскими учеными-психологами программа *Linguistic Inquiry and Word Count* [The Development and Psychometric Properties of LIWC2007] в версии для русского языка. Программа широко применяется зарубежными учеными для исследования проблемы взаимосвязи характеристик личности и параметров ее речевой продукции [Tausczik, Pennebaker, p. 24 – 54]. Она подсчитывает в тексте доли слов некоторых грамматических категорий, слов длиннее 6 букв, знаков препинания, слов тех или иных лексико-семантических категорий (слов, обозначающих положительные и отрицательные эмоции, восприятие, когнитивные процессы и т.д.), – всего 53 категории. Разработаны версии программы для английского, датского, китайского, арабского, французского, немецкого, итальянского, португальского, сербского, испанского, турецкого, русского языков. Отметим, что программа вычисляет не абсолютные, а относительные величины, что позволяет избежать зависимости от объема текста.

Многочисленные исследования с использованием данной программы (большинство из них выполнено на материале английского языка) подтвердили ее высокую эффективность для выявления языковых коррелятов как стабильных характеристик личности (пол, устойчивые психологические характеристики), так и различных состояний автора текста (стресса, болезни и т.д.) [Pennebaker, Mehl, Niederhoffer, p. 547 – 577; Pennebaker]. На материале русскоязычных текстов программа до настоящего времени не была апробирована.

Эксперимент проводили в два этапа. На первом этапе был проведен анализ 960 текстов от 480 авторов (по 2 текста от каждого автора), темы текстов – описание картины; письмо другу (назовем этот корпус условно K1). Для проверки полученных результатов на втором этапе работы нами были проанализированы 228 текстов от 114 авторов. Каждый из авторов писал 1 правдивый и 1 ложный текст на тему «Как я провел вчерашний день» (условно назовем этот корпус K2). Все проанализированные тексты имели примерно одинаковый объем, равный 130 – 160 словам.

Из параметров, которые можно вычислить при помощи LIWC, мы выбрали для последующих расчетов только те, значения которых превышали 1 % (т.е. низкочастотные параметры исключались из дальнейших расчетов). Отметим также, что мы использовали как обобщающие параметры (например, доля всех местоимений в тексте), так и более частные (доля личных местоимений в тексте).

Для определения степени стабильности признаков идиостиля мы использовали коэффициент вариации, который определяется по следующей формуле:

$$V = \frac{\sigma}{\bar{x}} \times 100\%,$$

где σ – среднее квадратическое отклонение, $\bar{x}$ – среднее арифметическое выборки [Левицкий].

Таким образом изначально, для определения устойчивых параметров текста мы рассчитали отклонение значений всех параметров от их средней величины по данным двух текстов одного и того же автора. Далее мы определили коэффициент вариации каждого параметра для всех авторов, что позволило нам дать абсолютную оценку меры разбросанности значений параметров текста и понять, насколько она велика относительно их среднего значения. Как и в наших предыдущих работах, посвященных диагностированию личности автора письменного текста на русском языке [Литвинова, 2013б, с. 64 – 72; Литвинова, Литвинова, Середин, с. 8 – 12], благодаря использованию в нашей работе статистического пакета SPSS мы получили возможность быстро и удобно обработать большой экспериментальный материал в исследовательских целях.

Результаты. Как показал статистический анализ, рассчитанный коэффициент вариации для выбранных параметров текста лежит в достаточно широких пределах. Среди исследуемых нами параметров имеются как относительно стабильные, так и отличающиеся значительной вариативностью. Принимая во внимание, что значение коэффициента вариации **менее 33 %** свидетельствует об однородности совокупности данных, в нашем случае такие значения коэффициента вариации означают относительную устойчивость параметра в текстах одного автора.

В группе устойчивых параметров нами были выделены две подгруппы: 1) подгруппа параметров, имеющих низкий коэффициент вариации (до 17 %); 2) подгруппа параметров с более высоким коэффициентом вариации (от 17 до 33 %).

Вслед за В.С. Андреевым назовем параметры первой подгруппы *линейно-стабильными*, второй подгруппы – *рекуррентно-стабильными* [Андреев].

К *подгруппе линейно-стабильных* в проанализированных нами текстах относятся следующие параметры (в скобках приведены соответствующие коэффициенты вариации для K1 и K2, в процентах): 1) доля слов длиннее 6 букв (15; 8); 2) доля строевых слов (11; 7); 3) доля предлогов (15; 10); 4) доля слов группы «Мыслительные процессы» (16; 13); 5) доля слов группы «Пространство» (15; 14); 6) доля знаков препинания (13; 8).

К *подгруппе рекуррентно-стабильных параметров* относятся следующие параметры: 1) доля местоимений (18; 17); 2) доля личных местоимений (21; 17); 3) доля союзов (26; 25); 4) доля слов группы «Эмоции» (26; 23); 5) доля слов группы «Положительные эмоции» (33; 26); 6) доля слов группы «Время» (31; 18).

К группе *нестабильных* относятся следующие параметры: 1) доля наречий (47; 36); 2) доля слов группы «Отрицание» (47; 47); 3) доля числительных (55; 59); 4) доля слов, обозначающих негативные эмоции (53; 55); 5) доля слов, входящих в отдельные подгруппы категории «Мысли-

тельные процессы»: «Объединение» (37; на K2 – 27), «Исключение» (41; 38); на K2 вычислялись и доли слов других подгрупп данной категории: «Интуиция» (34), «Причинно-следственные отношения» (40), «Несоответствие» (47), «Попытка» (47), «Уверенность» (37), «Сдерживание» (54), 6) доля слов группы «Восприятие» в целом (51; 37) и ее составляющих (исследовалось только на K2): подгрупп «Зрение» (45), «Слух» (47), «Ощущение» (46).

Обсуждение и выводы. Анализ результатов статистической обработки данных позволяет сделать следующие выводы.

1. Данные о стабильности/изменчивости, полученные на различных корпусах текстов, показывают хорошую сходимость, что, по нашему мнению, подтверждает правильность выбранного нами подхода. Возможно, более высокие (в среднем) значения коэффициента вариации на первом корпусе текстов (K1) обусловлены его большим объемом. Можно предположить, что с увеличением выборки увеличивается и коэффициент вариации, однако примечательно, что тем не менее полученные на разных корпусах данные согласуются, т. е. состав устойчивых и переменных параметров, выявленных на разных корпусах, в целом совпадает.

2. Выделение среди параметров идиостиля линейно- и рекурсивно-стабильных, с одной стороны, и переменных – с другой имеет важное значение как в теоретическом, так и в практическом отношении. Очевидно, что при проведении стилистических исследований необходимо учитывать тип вариативности того или иного параметра. Так, при решении задач атрибуции и диагностирования устойчивых характеристик личности особое внимание должно быть уделено стабильным признакам идиостиля. При решении вопросов, связанных с изменением идиостиля на протяжении жизни автора текста, а также с определением психического состояния человека в момент создания текста, с установлением факта наличия/отсутствия в тексте намеренно искаженной информации на первый план выходят переменные признаки.

3. Наше исследование, выводы которого, безусловно, являются предварительными, подтвердило правомерность использования компьютерных программ, и в частности программы LIWC, для обработки большого массива текстов с целью выделения стабильных и переменных характеристик идиостиля. Однако в дальнейшем необходима серьезная работа по расширению и уточнению состава входящих в нее словарей (как это было сделано, например, для китайской версии программы [Chinese LIWC Dictionary Development]).

По нашему мнению, особое внимание при этом следует уделить созданию и разработке словаря строевых слов ввиду крайней важности анализа данной группы лексики для построения методики идентификации и диагностирования личности автора текста. Отметим, что в нашей работе, выполненной на материале естественной письменной речи рядовых носителей языка, как и в работе [Фоменко, Фоменко, с. 688 – 722], выполненной, как мы уже отмечали, на материале художественных

произведений, было установлено, что доля служебных слов относится к числу высоко устойчивых параметров идиостиля.

4. Необходимо продолжать исследование проблемы стабильности/вариативности формализуемых параметров идиостиля, в том числе рядовых носителей языка. В нашей работе мы анализировали тексты большого числа респондентов, что позволяет говорить о достаточной репрезентативности полученных результатов, однако очевидно, что в дальнейших исследованиях необходимо увеличивать число текстов от одного респондента.

Помимо LIWC для анализа большого массива текстов и выявления зон стабильности и вариативности идиостиля также возможно использование синтаксических и морфологических анализаторов.

Подводя итог, следует отметить, что проблема устойчивости/изменчивости количественных параметров идиостиля далека от своего разрешения. Пока что в этой области больше вопросов, чем ответов: неясно, каково соотношение механизмов вариативности и стабильности идиостиля, каковы пределы варьирования параметров и каков масштаб их возможных колебаний, не изучен механизм взаимодействия параметров текста (варьируют ли признаки текста совместно либо же имеет место независимое варьирование признака и др.). Данные вопросы имеют не только теоретическую, но и практическую значимость: без ответов на них невозможно, в частности, построение по-настоящему валидных методик автороведческой экспертизы (как идентификационной, так и диагностической). Мы считаем, что для решения названных проблем необходимо активно привлекать существующие и разрабатывать новые компьютерные средства обработки языка, с помощью которых можно будет быстро и эффективно анализировать большие корпуса текстов.

Автор выражает глубокую благодарность доктору физико-математических наук П.В. Середину за помощь в статистической обработке материала

Литература

Андреев В.С. Языковая модель развития индивидуального стиля (на материале стихотворных текстов американских поэтов-романтиков): дис.... д-ра филол. наук. Смоленск, 2012.

Загоровская О.В., Литвинова Т.А., Литвинова О.А. Электронный корпус студенческих эссе на русском языке и его возможности для современных гуманитарных исследований // Мир науки, культуры и образования. 2012. № 3(34).

Левицкий В.В. Квантитативные методы в лингвистике. Черновцы, 2004.

Литвинова Т.А. Профилирование автора письменного текста // Язык и культура. 2013а. № 3 (23).

Литвинова Т.А. Формально-грамматические корреляты личностных особенностей автора письменного текста // Филологические науки. Вопросы теории и практики. 2013б. № 12 (30), ч. 1.

Литвинова Т.А., Литвинова О.А., Середин П.В. Частоты встречаемости последовательностей частей речи в тексте и психофизиологические характеристики его автора: корпусное исследование // Вестн. Иркутского гос. лингвистического ун-та. 2014. № 2.

Носенко Э.Л. Особенности речи в состоянии эмоциональной напряженности. Днепропетровск, 1975.

Фоменко В.П., Фоменко Т.Г. Авторский инвариант русских литературных текстов. Приложение. Кто был автором «Тихого Дона?» // Фоменко А. Т. Античность – это средневековье. СПб., 2005.

Chinese LIWC Dictionary Development [Электронный ресурс]. URL: <https://sites.google.com/site/taiwanliwc/english/chinese-liwc-dictionary-development> (дата обращения: 19.05.2015).

Koppel M., Argamon S., Shimoni A.R. Automatically categorizing written texts by author gender // Literary and Linguistic Computing. 2002. Vol. 17, № 4.

Koppel M., Schier J., Argamon Sh. Computational Methods in Authorship Attribution // Journal of the American Society for Information Science and Technology. 2008. Vol. 60. Iss. 1.

New machine learning methods demonstrate the theorized existence of a human stylome / H. van Halteren, R.H. Baayen, M. Haverkort [et al.] // Journal of Quantitative Linguistics. 2005. Vol. 12.

Pennebaker J. W. The Secret Life of Pronouns: What Our Words Say About Us. Bloomsbury Publishing USA, 2011.

Pennebaker J.W., Lay T.C. Language use and personality during crises: Analyses of Mayor Rudolph Giuliani's press conferences // Journal of Research in Personality. 2002. Vol. 36.

Pennebaker J.W., Mehl M.R., Niederhoffer K. Psychological aspects of natural language use: Our words, our selves. Annual Review of Psychology. 2003. Vol. 54.

Pennebaker J.W., Stone L.D. Words of wisdom: language use over the lifespan // Journal of Personality and Social Psychology. 2003. Vol. 85, Iss. 2.

Tausczik Y. R., Pennebaker J. W. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Method // Journal of Language and Social Psychology. 2010. Vol. 29, Iss. 1.

The Development and Psychometric Properties of LIWC2007 [Электронный ресурс] / Pennebaker J. W. [et al.] URL: http://homepage.psy.utexas.edu/homepage/faculty/pennebaker/reprints/liwc2007_languagemanual.pdf (дата обращения: 19.05.2015).

References

Andreev V.S. Yazykovaya model' razvitiya individual'nogo stilya (na materiale stikhotvornykh tekstov amerikanskikh poetov-romantikov): dis.... d-ra filol. nauk. Smolensk, 2012.

Zagorovskaya O.V., Litvinova T.A., Litvinova O.A. Elektronnyi korpus studencheskikh esse na russkom yazyke i ego vozmozhnosti dlya sovremennykh gumanitarnykh issledovaniy // Mir nauki, kul'tury i obrazovaniya. 2012. № 3(34).

Levitskii V.V. Kvantitativnye metody v lingvistike. Chernovtsy, 2004.

Litvinova T.A. Profilirovanie avtora pis'mennogo teksta // Yazyk i kul'tura. 2013a. № 3 (23).

Litvinova T.A. Formal'no-grammaticheskie korrelyaty lichnostnykh osobennostei avtora pis'mennogo teksta // *Filologicheskie nauki. Voprosy teorii i praktiki*. 2013b. № 12 (30), ch. 1.

Litvinova T.A., Litvinova O.A., Seredin P.V. Chastoty vstrechaemosti posledovatel'nostei chastei rechi v tekste i psikhofiziologicheskie kharakteristiki ego avtora: korpusnoe issledovanie // *Vestnik Irkutskogo gosudarstvennogo lingvisticheskogo universiteta*. 2014. № 2.

Nosenko E.L. Osobennosti rechi v sostoyanii emotsional'noi napryazhennosti. Dnepropetrovsk, 1975.

Fomenko V.P., Fomenko T.G. Avtorskii invariant russkikh literaturnykh tekстов. Prilozhenie. Kto byl avtorom «Tikhogo Dona?» // Fomenko A. T. Antichnost' – eto srednevekov'e. SPb., 2005. S. 688-722.

Chinese LIWC Dictionary Development [Elektronnyi resurs]. URL: <https://sites.google.com/site/taiwanliwc/english/chinese-liwc-dictionary-development> (data obrashcheniya: 19.05.2015).

Koppel M., Argamon S., Shimoni A.R. Automatically categorizing writ-ten texts by author gender // *Literary and Linguistic Computing*. 2002. Vol. 17, No. 4.

Koppel M., Schier J., Argamon Sh. Computational Methods in Au-thorship Attribution // *Journal of the American Society for Information Science and Technology*. 2008. Vol. 60. Iss. 1.

New machine learning methods demonstrate the theorized existence of a human stylome / H. van Halteren, R.H. Baayen, M. Haverkort [et al.] // *Journal of Quantitative Linguistics*. 2005. Vol. 12.

Pennebaker J. W. *The Secret Life of Pronouns: What Our Words Say About Us*. Bloomsbury Publishing USA, 2011.

Pennebaker J.W., Lay T.C. Language use and personality during crises: Analyses of Mayor Rudolph Giuliani's press conferences // *Journal of Research in Personality*. 2002. Vol. 36.

Pennebaker J.W., Mehl M.R., Niederhoffer K. Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*. 2003. Vol. 54.

Pennebaker J.W., Stone L.D. Words of wisdom: language use over the lifespan // *Journal of Personality and Social Psychology*. 2003. Vol. 85, Iss. 2.

Tausczik Y. R., Pennebaker J. W. The Psychological Meaning of Words:LIWC and Computerized Text Analysis Method // *Journal of Language and Social Psychology*. 2010. Vol. 29, Iss. 1.

The Development and Psychometric Properties of LIWC2007 [Elektronnyi resurs] / Pennebaker J. W. [et al.] URL: http://homepage.psy.utexas.edu/homepage/faculty/pennebaker/reprints/liwc2007_languagemanual.pdf (data obrashcheniya: 19.05.2015).

Litvinova T.A. (Voronezh state pedagogical university, Voronezh, Russian Federation)

To the problem of idiostyle characteristics stability

The problem of stability and variability of author's idiostyle refers to the number of less studied problems. In some works it is decided mainly on the material of works of art. The article presents the results of the pilot studies, of the stability/variability of quantitative parameters of the texts, which are samples of Russian natural writing. The texts were processed using the Russian version of the program Linguistic Inquiry and Word

Count, developed by American scientists-psychologists. This program has been used successfully to study the problem of the relationship of language and identity on the material of a number of languages (English, German, Chinese, Italian, Arabic and others.), but in Russian material it has not been studied. The program estimates the proportion of words in the text of certain morphological and lexical-semantic categories. 1188 texts were analyzed by 594 respondents, and the methods of statistical analysis obtained coefficients of variation for each language parameters. It is highlighted text parameters being relatively stable in texts of one author, and the parameters that have high coefficients of variation-coefficient. The possibility of using the results of IC solutions for the following tasks of author study and prospects for further work are defined.

Key words: *stylemetrics, author study, idiostyle, written speech, attribution of the text, corpus.*

Litvinova Tatyana Alexandrovna – Voronezh state pedagogical university, Regional center of the Russian language, candidate of philology, research worker.

Phone: 8-980-342-00-73, e-mail: centr_rus_yaz@mail.ru