Ковалев А.В.

ПРОИСХОЖДЕНИЕ ТРАКТАТА “DESCRIPTIO EUROPAE ORIENTALIS”: К ВОПРОСУ О ПРИМЕНЕНИИ МЕТОДИК КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ ДЛЯ АНАЛИЗА СРЕДНЕВЕКОВЫХ ТЕКСТОВ[*]

Всякий средневековый текст, пусть и написанный на латыни, не был создан носителям латинского языка. Исходя из этого суждения, мы можем говорить, что латынь всех средневековых текстов в той или иной степени является искаженной сквозь призму родного языка автора, принадлежащего к другой языковой культуре. Поэтому в средневековых текстах практически всегда присутствуют грамматические и орфографические элементы, несвойственные классической латыни. В то же время, по причине того, что язык является системой, эти изменения также не могут быть бессистемными и большая часть девиаций латинского языка в средневековых текстах может быть объяснена лингвистически.

Значительная часть нестандартных оборотов в таких текстах может быть следствием наложения на латинский текст грамматических и орфографических систем, свойственные либо родному языка автора – особенно если родной язык создателя текста принадлежал к романской группе – либо вообще представляющих из себя наложение нескольких романских языков на латинскую первооснову, в случае если автор владел другими языками помимо латыни и своего родного.

Обозначенная особенность средневековых текстов может быть достаточно полезна в деле локализации и в ряде случаев атрибуции средневековых письменных памятников. На практике данный подход к средневековым текстам позволяет по-новому взглянуть на источники, до текущего момента остававшиеся анонимными, а так же на сочинения, чьи авторы были установлены с неабсолютной достоверностью.

Целью данного доклада является обращение к одному из таких письменных памятников, анонимному средневековому трактату “Descriptio Europae Orientalis”[1]. Этот трактат был создан по заказу Папской курии в 1307 – начале 1308 года и был призван заполнить лакуну в знаниях крестоносцев о Восточноевропейском географическом пространстве. Трактат был включен в качестве одного из составных элементов в папскую программу крестового похода, конечной целью которого было отвоевание Иерусалима и восстановление господства франков в Леванте.

Уже при беглом взгляде на текст “Descriptio Europae Orientalis” обращают на себя внимание такие характерные особенности текста как характерные для средневековой латыни перестановки “ti” –“ci”, “v”–“u”, замена дифтонга “ae”” на e”, “j”–“i”, etc.

Кроме того, в процессе перевода текста источника был обнаружен целый ряд слов, не специфичных для классической латыни. Часть из них оказалась словами греческого происхождения, которые иногда встречаются в средневековых латинских письменных сочинениях. К таковым относятся, например, слова: Exeniat, от греческого ξένος – чужой, дар, подношение гостю; Calogeri, от греческого καλόγερος – монах, etc.

Однако наиболее интересен для нас блок слов романского происхождения, но не характерных для классической латыни. Данный блок лексики был сгруппирован в таблице 1. Мы можем говорить, что грамматические категории слов в первом столбце таблицы: имя или глагол, а также род число, падеж лицо и т.д. реконструируются достаточно легко, в силу того, что эта лексика согласована по правилам латинской грамматики и представлена в связанном и несущем смысловую нагрузку тексте

[*] This publication is a part of my research work at Lund University, thanks to a Swedish Institute scholarship and the Historical Department of Lund University.

Репозиторий ВГУ

“Descriptio Europae Orientalis”. Поэтому для слов в первом столбце, опираясь на данные грамматики и семантики предложений, были подобраны эквиваленты в итальянском и французском языках, которые являются наиболее вероятными родными языками для автора исследуемого трактата. Кроме того, для увеличения точности исследования в таблицу был введен дополнительным элементом, третий романский язык – испанский.

Таблица 1

DEO	Italiano	Frances	Español
amena – приятная	*amena*	-	*amena*
apostatauit – отреклись	*apostatare*	-	-
deliciosa – приятная	*deliziosa*	*délicieux*	*deliciosa*
eque – равные	*eguali*	*egale*	-
foratto – подбитый мехом	*foderato*	*fuorré*	*forrado*
foratto – украшенный цветами	*fiorato*	*fleuri*	*florido*
husonibus – дунайский лосось	*huco*	*hucho*	*hucho*
natural – натуральный	*naturale*	*naturel*	*natural*
roncini – клячи	*ronzini*	*rosses*	*rocines*
sarcita – наполненная	*farcita*	*farci*	-
scarlatto – ярко красный	*scarlatto*	*scarlatine*	-
stutionibus – осетр	*storione*	*esturgeon*	*esturión*

Для анализа собранных данных, можно использовать алгоритм под названием Дистанция Левенштейна. Решение подобной проблемы было впервые было упомянуто в 1965 году советским математиком В.И. Левенштейном [2], в честь которого эта задача и была впоследствии названа. Сегодня этот алгоритм активно используется в компьютерной лингвистике для решения задач коррекции орфографических ошибок в текстовых редакторах. Дистанция Левенштейна, на западе также известная как Минимальная дистанция редактирования (Minimum edit distance)–это система измерения разницы между двумя последовательностями (в данном случае словами), т.е. минимально необходимое количество изменений одиночных знаков последовательности (в данном случае букв): удаление, вставка, замена, для превращения одной последовательности в другую.

В виде формулы дистанция Левенштейна может быть представлена следующим образом:

Для 2-х последовательностей (слов):

X, с длинной = n

Y, с длинной = m

D(istance) (i,j) = x[1...*i*] и y[1...*j*]

Инициализация: D(*i*, 0)=*i*

D(0,*j*)=*j*

$$D(i,j) = \min \begin{cases} D(i-1,j) + 1 & \text{(Удаление)} \\ D(i, j-1) + 1 & \text{(Вставка)} \\ D(i-1,j-1) \begin{cases} +2 & \text{if } x(i) \neq y(j) \\ +0 & \text{if } x(i) = y(j) \end{cases} & \text{(Замена)} \\ D(i-1,j-1) + 0, \text{ if } x(i, i-1) = y(j, j-1) & \text{(Перестановка)} \end{cases}$$

Репозиторий БГУ

В качестве иллюстрации для анализа текста в рамках данного доклада нами было выбрано слово “forrato”. Оно встречается в тексте “Descriptio Europae Orientalis” всего один раз в сообщении об одежде греческой знати:

Omnes principes grecie, ac ceteri nobiles et omnes de imperatoris familia uadunt induti sericeis et deauratis pannis uel scarleto *forrato* de nobilibus pellibus [1, p, 22].

Особенностью данного слова является тот факт, что, опираясь только на сходство с лексикой других романских языков, не возможно достоверно определить семантическое значение данной лексической единицы.

Так “forrato”может быть интерпретировано либо как итальянское “fiorato” – украшенный узорами, цветами, либо как итальянское – “foderato” подбитый мехом, либо же, как испанское “florido” – цветочный, либо как испанское “forrado” – подбитый мехом.

Ниже представлены матрицы по расчету минимальной дистанции редактирования для всех 4 вероятных слов-прародителей:

Таблица 2

Fiorato - украшенный узорами

O	**7**	6	7	6	5	4	3	**2**
T	**6**	5	6	5	4	3	2	3
A	**5**	4	4	4	3	2	3	4
R	**4**	3	4	3	2	3	4	5
R	**3**	2	3	2	1	2	3	4
O	**2**	1	2	1	2	3	4	5
F	**1**	0	1	2	3	4	5	6
#	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
	#	**F**	**I**	**O**	**R**	**A**	**T**	**O**

Foderato - подбитый мехом

O	**7**	6	5	6	7	6	5	4	**3**
T	**6**	5	4	5	6	5	4	3	4
A	**5**	4	3	4	5	4	3	4	5
R	**4**	3	2	3	4	3	4	5	6
R	**3**	2	1	2	3	2	3	4	5
O	**2**	1	0	1	2	3	4	5	6
F	**1**	0	1	2	3	4	5	6	7
#	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
	#	**F**	**O**	**D**	**E**	**R**	**A**	**T**	**O**

Florido - украшенный узорами

O	**7**	6	7	6	5	6	7	**6**
T	**6**	5	6	5	4	5	6	7
A	**5**	4	5	4	3	4	5	6
R	**4**	3	4	3	2	3	4	5
R	**3**	2	3	2	1	2	3	4
O	**2**	1	2	1	2	3	4	5
F	**1**	0	1	2	3	4	5	6
#	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
	#	**F**	**L**	**O**	**R**	**I**	**D**	**O**

Репозиторий ВГУ

Forrado - подбитый мехом

O	**7**	6	5	4	3	2	3	**2**
T	**6**	5	4	3	2	1	2	3
A	**5**	4	3	2	1	0	1	2
R	**4**	3	2	1	0	1	2	3
R	**3**	2	1	0	1	2	3	4
O	**2**	1	0	1	2	3	4	5
F	**1**	0	1	2	3	4	5	6
#	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
	#	**F**	**O**	**R**	**R**	**A**	**D**	**O**

Сводные результаты проведенной обработки для всех данных из таблицы 1 представлены в таблице 3. Прочерк в таблице 3 обозначает либо отсутствие слова для сравнения в языке, представленном в колонке, либо же слишком большую величину дистанции по результатам подсчетов, что также позволяет не учитывать такие данные.

Таблица 3

DEO	Italiano	Frances	Español
Amena	0	-	0
Apostatauit	0	-	-
Deliciosa	2	-	0
Eque	6	5	-
Forrato no1	2	-	6
Forrato no2	3	-	2
Husonibus	5	6	6
Natural	1	2	0
Roncini	2	-	4
Sarcita	0	2	-
Scarleto	3	5	-
Stutionibus	5	8	5
Среднее	**2,41**	**4,57**	**3,16**

На основании полученных данных можно утверждать следующее:

- Хотя в ряде случаев (например, “forrato” – “forrado”) дистанция Левенштейна является минимальной для испанского языка, мы не можем считать, что лексика трактата имеет испанское происхождение. Это связано с тем фактом, что в остальных случаях дистанция редактирования для испанского языка чрезвычайно велика, а ряду слов вообще невозможно найти подходящий по семантике и орфографии эквивалент. Средняя дистанция 3,16.
- В случае с анализом французского сегмента лексики, мы можем констатировать, что отличия нетипичной лексики от французского языка еще большие, чем от испанского языка. Средняя дистанция составляет 4,57 и это без учета слов с дистанцией больше 10. Все это ставит под сомнение гипотезу, в соответствии с которой возможным автором трактата был француз, папский секретарь Николя Фалько из Тула (Nicolas Falcon de Toul).

Репозиторий ВГУ

- Наименьшая дистанция 2,41, а также наличие эквивалентов ко всем словам позволяют предположить со значительной долей уверенности, что лексика трактата имеет итальянское происхождение.

1. Anonymi descriptio Europae orientalis. Imperium Constantinopolitanum, Albania, Serbia, Bulgaria, Ruthenia, Ungaria, Polonia, Bohemia. / ed. O. Górka. – Cracoviae: Sumptibus Academiae Litterarum, 1916. – 70 p.
2. Левенштейн, В.И. Двоичные коды с исправлением выпадений, вставок и замещений символов / В.И. Левенштейн // Доклады Академий Наук СССР. – М.,1965. – С. 845–848.

Репозиторий ВГУ