М. Е. Васильев[†], *М. Н. Саенко*[‡]
[†] Школа актуальных гуманитарных исследований РАНХиГС (Россия, Москва), mvhumanity@gmail.com;
[‡] Институт славяноведения РАН (Россия, Москва), veraetatis@yandex.ru

К вопросу о точности глоттохронологии: датирование процесса лексических замен по данным романских языков

Статья представляет собой первую часть исследования, посвященного проблеме достоверности лингвистических датировок, получаемых с помощью метода глоттохронологии. В предлагаемой работе рассматривается процесс лексических замен, происходящих в базисной лексике одного языка с течением времени. В качестве исходных данных нами использовались 110-словные списки, собранные на материале 54 современных и нескольких исторических романских идиомов. При этом для измерения скорости замен списки современных языков сравнивались со списками классической и архаический латыни, а также старофранцузского и староитальянского. Временна́я дистанция между сопоставляемыми идиомами определялась с помощью трех различных глоттохронологических методов: классического уравнения М. Сводеша, модифицированной формулы С. А. Старостина, а также недавно предложенной потоковой модели. Сравнение полученных результатов позволило сделать ряд важных выводов о характере лексических изменений, адекватности существующих глоттохронологических моделей, а также численно оценить точность и надежность глоттохронологических расчетов при датировании общего процесса замен. Вторую часть исследования планируется посвятить проблеме датирования относительной дивергенции двух родственных языков.

Ключевые слова: глоттохронология, лексикостатистика, список Сводеша, романские языки.

Одной из основных аксиом лексикостатистики является равномерность процесса замен базисной лексики, описанная М. В. Араповым и М. М. Херц следующим образом:

> *Доля p* слов из О[сновного] С[писка], которые сохранятся (не будут заменены другими словами) на протяжении интервала времени Δt (равного, например, году, столетию или тысячелетию) постоянна (т. е. зависит только от величины выбранного промежутка, но не от того, как он выбран, или слова какого языка рассматриваются). (Арапов, Херц 1974: 22)

Математическим соответствием этой аксиомы является коэффициент сохраняемости, используемый в глоттохронологических формулах вычисления времени. Впервые коэффициент сохраняемости был вычислен М. Сводешем путем сравнения двухсотсловных списков ряда языков с долгой письменной историей в ранней форме их существования и на более позднем этапе развития: табл. 1.

Как можно заметить, процент сохранившейся лексики при данной методике подсчета колеблется в интервале 76—85 % слов за тысячелетие (Сводеш 1960a: 47).

В дальнейшем, используя модифицированный список базисной лексики, состоящий из 100 слов, Сводеш пересчитывает на его основании коэффициент сохраняемости, предварительно исключив из рассмотрения коптский (так как он может быть непрямым

Таблица 1. Значения коэффициента сохраняемости, рассчитанные Сводешем на основании двухсотсловных списков (Сводеш 1960а: 34).

Ранняя форма	Поздняя форма	Интервал времени	Процент сохранившихся слов на 1000 лет
Среднеегипетский	Коптский	2300 лет	76
Классическая латынь	Румынский	2000 лет	77
Древневерхненемецкий	Немецкий	1100 лет	78
Классический китайский	Северокитайский	1000 лет	79
Латынь Плавта	Французский Мольера	1850 лет	79
Доминика кариб 1650 н.э.	Современная форма	300 лет	80
Классическая латынь	Португальский	2000 лет	82
Койне	Кипрский диалект	[без даты]	83
Койне	Афинский диалект	[без даты]	84
Классическая латынь	Итальянский	2000 лет	85
Древнеанглийский	Английский	1000 лет	85
Латынь Плавта	Ранний новоиспанский 1600 г.	1800 лет	85

Таблица 2. Значения коэффициента сохраняемости, рассчитанные Сводешем на основании стословных списков (Сводеш 1960б: 72).

Язык	Интервал времени	Процент сохранившихся слов на 1000 лет
Шведский	1020	94,3
Немецкий	1100	89,0
Английский	1000	76,6
Румынский	2150	76,4
Французский	1850	77,6
Афинский	2070	83,6
Китайский	1000	79,6

потомком среднеегипетского), испанский, итальянский, португальский, каталанский (поскольку уже есть подсчет по родственному ему французскому), кипрский (поскольку есть подсчет по афинскому): табл. 2.

Однако уже в 1958 г. появилась статья Дж. Ри, в которой сравниваются стословные списки восьми романских языков и при использовании коэффициента сохраняемости r = 0,85 (максимальный среди вычисленных Сводешем) были получены нелепые даты расхождения, например, испанского и португальского — 370 лет назад, румынского и итальянского — 826 лет назад. Поскольку история романских языков довольно хорошо известна, очевидно, что эти результаты не отвечают действительности. Исходя из этого, Дж. Ри делает вывод о некорректности всей методики глоттохронологии. В ответной статье А. Крубер возражает, что в данном случае нужно не сходу отказываться от глоттохронологии вообще, а попытаться ее доработать. В частности, для романских языков

Таблица 3. Расчеты К. Бергсланда и Х. Фогта (Bergsland, Vogt 1962: 117—125).

Язык-1	Язык-2	Временно́е расстояние (100-словник)	Временно́е расстояние (200- словник)	Временно́е расстояние (215- словник)	Фактическая временна́я дистанция
Древне-скандинавский	Исландский	63	130	194	1000
Древне-скандинавский	диалект Гьестал	345	799	901	1000
Древне-скандинавский	диалект Санднес	412	861	964	1000
Древне-скандинавский	Риксмол	637	930	1000	1000
Древне-грузинский	Грузинский	338	750	861	1500
Грузинский	Мегрельский	1316	1004	1033	1800—1900
Древне-армянский	Армянский	211	437	437	1500

Таблица 4. Значения коэффициента сохраняемости, полученные Старостиным на материале стословных списков (Starostin 2000: 230).

Язык	Интервал времени	λ_1 (с учетом заимствований)	λ_2 (без учета заимствований)
Японский	1200 лет	0,11	0,06
Китайский	2600 лет	0,1	0,1
Английский	1300 лет	0,14	0,1
Немецкий	1200 лет	0,08	0,05
Французский	1500 лет	0,09	0,07
Испанский	1500 лет	0,07	0,06
Румынский	1500 лет	0,09	0,06

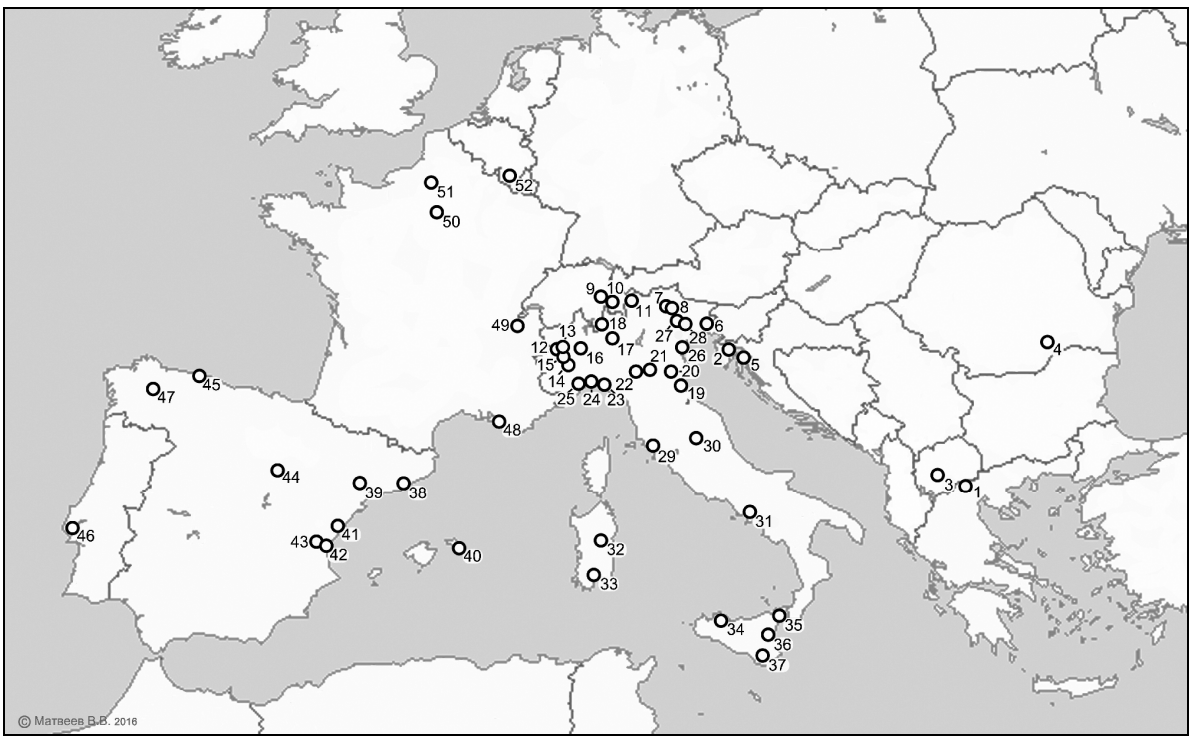
при введении поправочного коэффициента 2,2 Крубером были получены вполне приемлемые даты (Kroeber 1958).

Тем не менее, последующие подсчеты, произведенные по методике Сводеша К. Бергсландом и Х. Фогтом (Bergsland, Vogt 1962) на материале риксмола, исландского, армянского и грузинского языков, снова дали результат, существенно расходящийся с известными значениями: табл. 3.

Как видно из табл. 3, кроме единственного исключения, расчетные значения временной дистанции ни в одном случае не совпадают с фактическими. Особенно разительным оказалось расхождение при сравнении исландского и древнескандинавского языков.

В середине 80-х годов прошлого века усовершенствованием метода глоттохронологии занялся С. А. Старостин, который улучшил формулу дивергенции языков и выдвинул требование исключать заимствования из списков сопоставляемой лексики, поскольку заимствования являются следствием внешних контактов, а не внутреннего изменения языка: табл. 4.

Рисунок 1. Географическое распространение идиомов, используемых в исследовании. Номерами на карте обозначены: 1. мегленорумынский; 2. истрорумынский; 3. арумынский; 4. румынский (литературный); 5. далматинский; 6. фриульский (центральный); 7. ладинский (гарденский); 8. ладинский (фассанский); 9. руманшский (сурсельвский); 10. руманшский (сурмиранский); 11. руманшский (нижнеэнгадинский); 12. пьемонтский (*Ланцо-Торинезе*); 13. пьемонтский (Барбания); 14. пьемонтский (Карманьола); 15. пьемонтский (Турин); 16. пьемонтский (верчельский); 17. ломбардский (Бергамо); 18. ломбардский (Плезио); 19. эмилиано-романьольский (Равенна); 20. эмилиано-романьольский (Феррера); 21. эмилиано-романьольский (Карпи); 22. эмилиано-романьольский (Реджо); 23. лигурийский (Рапалло); 24. лигурийский (Генуя); 25. лигурийский (Стелла); 26. венетский (Венеция); 27. венетский (Примьеро); 28. венетский (Беллуно); 29. тосканский (Гроссето); 30. умбрийский (Фолиньо); 31. неаполитанский; 32. логудорский; 33. кампиданский; 34. сицилийский (Палермо); 35. сицилийский (Мессина); 36. сицилийский (Катания); 37. сицилийский (юго-восточный; объединены данные, полученные от информантов из Рагузы и Агридженто); 38. каталанский (центральный); 39. каталанский (северо-западный); 40. каталанский (Менорка); 41. каталанский (Кастельон-де-ла-Плана); 42. каталанский (Валенсия); 43. каталанский (Манисес); 44. кастильский (Сория); 45. астурийский (центральный); 46. португальский (литературный); 47. галисийский (центральный); 48. окситанский (провансальский); 49. франко-провансальский (савойский); 50. французский (литературный); 51. пикардский (южный); 52. валлонский (южный).



За последние годы в рамках проекта «Глобальная лексикостатистическая база данных» было накоплено большое количество списков базисной лексики, качество которых стоит на более высоком уровне, чем у материала, доступного Сводешу и Старостину.

В частности, в 2015—2016 гг. одним из авторов данной статьи (М. Н. Саенко) были собраны аннотированные 110-словные списки базисной лексики для 54 [1] романских идиомов (а также 4 списка для староитальянского Данте, старофранцузского Кретьена де Труа, ла-

[1] В дальнейшем будут использоваться 52 идиома, так как списки для руманч грижун и итальянского литературного были исключены в силу искусственного характера первого и чрезвычайной архаичности второго. Полные списки с источниками, таблицами транслитерации и описанием доступны на сайте «Глобальная лексикостатистическая база данных» http://starling.rinet.ru/cgi-bin/main.cgi?root=new100.

тыни Плавта и Апулея). При этом предпочтение отдавалось диалектам и «малым» языкам без строгой литературной нормы, из 54 идиомов лишь 5 являются строго нормированными (румынский, руманч грижун, итальянский, португальский, французский). Географически исследование было ограничено только Старым Светом (см. рис. 1). Данные были получены как из диалектных словарей романских языков, так и от информантов. В связи с необходимостью за короткий срок охватить большое количество материала, работа с носителями велась через Интернет (рассылался опросник, а после его заполнения информантам задавались уточняющие вопросы). Всего было опрошено 76 информантов.

Используя материал романских языков в качестве тестового, мы последовательно применяем к нему два наиболее известных глоттохронологических метода: классическую глоттохронологию М. Сводеша, усовершенствованную методику С. А. Старостина, а также недавно предложенную потоковую модель. При этом определим основные цели и задачи нашего исследования следующим образом:

1) Проверить применимость существующих глоттохронологических моделей (М. Сводеша, С. А. Старостина, потоковой модели) для датирования изменений в базисной лексике романских языков;
2) Определить оптимальные параметры моделей, обеспечивающие наилучшее соответствие расчетных значений и исходных лексикостатистических данных. При необходимости произвести калибровку моделей с учетом новых параметров и сравнить результаты, полученные с использованием калиброванных и некалиброванных моделей;
3) Установить объективные и теоретические пределы точности при вычислении лингвистических датировок с применением рассматриваемых моделей.

В первой части данной статьи мы рассмотрим процесс замен в лексике одного языка по мере его развития и подробно остановимся на методике вычисления временной дистанции между языком-предком и языком-потомком. Вторая часть работы будет посвящена относительной дивергенции двух родственных языков и проблемам ее датирования.

1. Глоттохронологические модели общего процесса лексических замен

Исходные данные для определения скорости изменений в базисной лексике романских языков были получены путем сравнения современных идиомов с ближайшими родственниками их непосредственных предков: архаической латынью Плавта, поздней классической латынью Апулея, староитальянским Данте и старофранцузским Кретьена де Труа, датировки которых можно установить по историческим источникам. По результатам сравнения составлена табл. 5, где для каждой пары (или нескольких пар)[2] языков указан процент совпадений[3] между соответствующими основными списками, а также временной интервал между датами их фиксации.

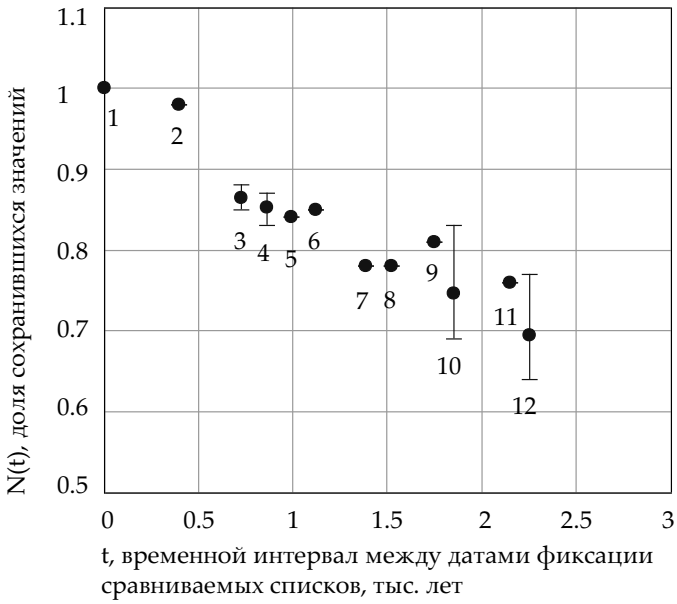
[2] При сопоставлении нескольких пар языков (строки 3, 4, 10, 12), в соответствующих столбцах приводится минимальное, максимальное и среднее арифметическое значение. Последнее рассчитывается как сумма всех известных процентов совпадений, деленная на количество слагаемых, что в общем случае не совпадает со средним значением между максимальным и минимальным процентом совпадений. Например, в строке 4 для трех пар языков с долями совпадений 86%, 83% и 87% *среднее* между минимальным и максимальным значением составит (83+87)/2=85, а *среднее арифметическое* — (83+86+87)/3=85,3.

[3] Значения процентов совпадений приводятся по данным табл. 8, полученной путем попарного сравнения всех собранных списков в программе Starling (см. Дополнительные материалы).

Таблица 5. Лексикостатистические данные о скорости замен в базисной лексике романских языков.

№	Сравниваемые языки	Мин. % совп.	Средн. % совп.	Макс. % совп.	Интервал времени, лет
1	Исходное значение (для любого идиома)	—	100	—	0
2	Архаическая латынь (Плавт, 250 г. до н.э.) — поздняя классическая латынь (Апулей, 150 г. н.э.)	—	98	—	400
3	Староитальянский (Данте, 1270 г.) — современные итальянские (тосканский, умбрийский)	85	86,5	88	730
4	Старофранцузский (Кретьен де Труа, 1140 г.) — современные французские (литературный французский, пикардский, валлонский)	83	85,3	87	860
5	Поздняя классическая латынь (Апулей, 150 г.) — старофранцузский (1140 г.)	—	84	—	990
6	Поздняя классическая латынь (150 г.) — староитальянский (1270 г.)	—	85	—	1120
7	Архаическая латынь (250 г. до н.э.) — старофранцузский (1140 г.)	—	78	—	1390
8	Архаическая латынь (250 г. до н.э.) — староитальянский (1270 г.)	—	78	—	1520
9	Поздняя классическая латынь (150 г.) — далматинский (1900 г.)	—	81	—	1750
10	Поздняя классическая латынь (150 г.) — современные романские (52 идиома, 2000 г.)	69	74,6	83	1850
11	Архаическая латынь (250 г. до н.э.) — далматинский (1900 г.)	—	76	—	2150
12	Архаическая латынь (250 г. до н.э.) — современные романские (52 идиома, 2000 г.)	64	69,5	77	2250

Рисунок 2. Изменение доли сохранившейся лексики романских языков в зависимости от времени. Нумерация рядом с точками указывает на соответствующие строки табл. 5. Для точек 3, 4, 10 и 12 показан диапазон разброса долей совпадений и среднеарифметическое значение.



Для наглядности полученные данные можно представить в виде диаграммы, которая отражает уменьшение доли сохранившейся лексики N(t) с течением времени (t): рис. 2.

Приведенная диаграмма свидетельствует о ярко выраженном статистическом характере исходных данных, который проявляется в значительном разбросе процентов совпадений, полученных для точек с одинаковыми или хронологически близкими датировками. В то же время, очевидно, что все (даже наиболее выделяющиеся) значения хорошо сгруппированы вокруг некоторой средней величины на всем рассматриваемом интервале времени, что позволяет говорить о существовании зависимости случайного процесса лексических замен от времени. Для определения характера этой зависимости и ее параметров перейдем к рассмотрению конкретных глоттохронологических моделей: классическому уравнению М. Сводеша, усовершенствованной формуле С. А. Старостина и потоковой модели.

1.1. Глоттохронологическая модель М. Сводеша.

Классический метод глоттохронологии, предложенный М. Сводешем в середине XX в., построен по аналогии с методом радиоуглеродного датирования и базируется на четырех основных допущениях (постулатах[4]):

а) наличие в словаре каждого языка некоторого устойчивого подмножества слов — базисной лексики, из которой можно выделить универсальный список значений, обладающий повышенной стабильностью в любом языке[5];
б) постоянная скорость лексических изменений в основном списке, не зависящая от выбранного языка и временного периода;
в) одинаковая стабильность всех элементов основного списка;
г) независимость замен в списках языков-потомков после их разделения.

В качестве математического аппарата, отражающего содержание постулатов глоттохронологии, была использована формула радиоактивного распада, описывающая процесс замен в базисной лексике в виде экспоненциальной зависимости с коэффициентом сохраняемости λ, определяющим темп замен:

$$N(t) = e^{-\lambda \cdot t}.$$

Как уже говорилось, согласно подсчетам Сводеша, которые проводились на разнообразном материале (в том числе романском), за 1000 лет различные языки в среднем сохраняют около 85% основного списка (Swadesh 1952: 456—460), что соответствует коэффициенту λ=0,16 ($e^{-0,16*1}$=0,852). Подставив данное значение λ в исходное выражение, получаем формулу для датирования процесса замен в лексике одного языка:

$$N_{Sw}(t) = e^{-0,16 \cdot t}.$$

Для калибровки формулы Сводеша по новым данным, полученным на основе романских языков, воспользуемся методом наименьших квадратов. Смысл данного метода сводится к поиску таких параметров модели (в данном случае — коэффициента λ), при

[4] Эти и другие положения глоттохронологии Сводеша более подробно излагаются в работе Арапов, Херц 1974: 21—25.

[5] Наибольшее распространение получил 100-словный список, зачастую называемый также «списком Сводеша».

Рисунок 3. Соответствие исходной и калиброванной модели М. Сводеша исходным данным: $N_{Sw}(t) = e^{-0,16 \cdot t}$; $N_{SwC}(t) = e^{-0,16 \cdot t}$.

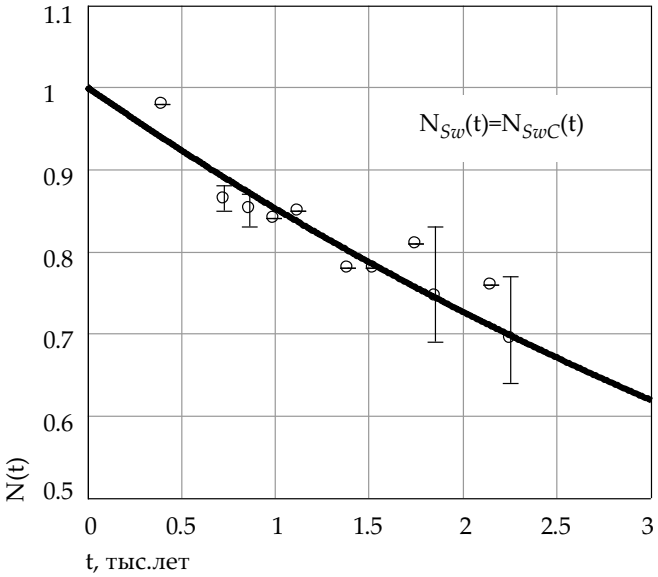
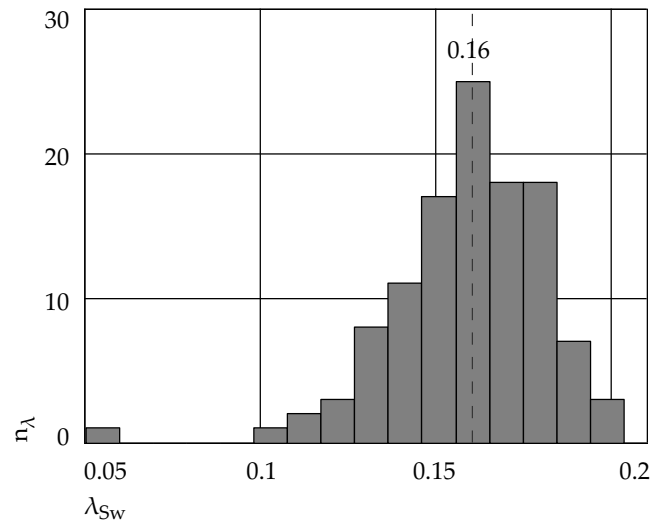


Рисунок 4. Количество пар языков (n_λ) с коэффициентом λ, полученным для калиброванной модели Сводеша по исходным данным (табл. 5).



которых суммарная разница (или отклонение) между фактическими и расчетными долями совпадений, вычисленными для каждой из опорных точек, является минимальной.

Например, в соответствии с табл. 5 (строка 2), процент совпадений между списками арахаической латыни Плавта и классической латыни Апулея составляет 98 (N_ϕ=0,98) при разделяющем их временном интервале 400 лет (t=0,4). Подставляя это значение времени в формулу Сводеша, получим выражение: $N_p = e^{-\lambda \cdot 0,4}$. Теперь, для того чтобы обеспечить наилучшее совпадение фактического (N_ϕ) и расчетного (N_p) значений, необходимо найти такую величину λ, при которой квадрат разности между фактическим и расчет-

ным значениями $(N_{ф} - N_{p})^2$ окажется наименьшим. Таким образом, задача поиска минимального отклонения ε для данной опорной точки будет иметь вид:

$$\varepsilon = (e^{-\lambda_1 \cdot 0,4} - 0,98)^2 \rightarrow \min.$$

Суммируя отклонения, полученные аналогичным способом для каждой опорной точки, получим общую формулу для расчета оптимального коэффициента λ:

$$\varepsilon = \sum_{i} (N_p - N_ф)^2 \rightarrow \min,$$

где N_p — расчетное значение доли общей лексики, вычисленное по формуле $N_{Sw}(t) = e^{-\lambda \cdot t}$, i — номер опорной точки, а $N_ф$ и t — фактические значения доли совпадений и времени, представленные в табл. 5.

Вычисления, проведенные с помощью пакета Mathcad, показали, что минимальное отклонение между расчетными и фактическими значениями достигается при коэффициенте сохраняемости равном 0,16 (при этом величина суммарного отклонения составляет ε=0,094).

Таким образом, калиброванное значение λ совпало с константой Сводеша, а полученная модель оказалась идентична исходной (см. рис. 3):

$$N_{SwC}(t) = N_{Sw}(t) = e^{-0,16 \cdot t}.$$

Как видно на приведенном графике (рис. 3), расчетные значения модели хорошо соответствуют опорным точкам на всем временном интервале (до 2,5 тыс. лет). При этом распределение значений λ, полученное для всех опорных точек по формуле Сводеша, оказалось близким к нормальному распределению[6] с математическим ожиданием $\bar{\lambda}_{Sw}$=0,16 и средним квадратическим отклонением $\sigma_\lambda \approx 0,02$ (рис. 4).

Воспользуемся теперь другой глоттохронологической моделью, которая была предложена С. А. Старостиным.

1.2. Усовершенствованная глоттохронология С. А. Старостина.

Анализируя критику методики Сводеша, С. А. Старостин приходит к выводу о несостоятельности 2-го и 3-го постулатов глоттохронологии и указывает на необходимость их пересмотра. В частности, он приводит следующие аргументы (Starostin 2000: 229—230, 236—237):

а) слова в базисной лексике языка со временем устаревают, и чем больше рассматриваемый промежуток времени, тем больше вероятность замены слова в основном списке, а следовательно — тем выше скорость распада.

б) слова в основном списке неоднородны и обладают разной стабильностью, поэтому с течением времени общая скорость распада снижается из-за повторных замен наименее устойчивых значений и увеличивающейся доли более устойчивых.

Чтобы учесть эти особенности в математическом аппарате глоттохронологии, С. А. Старостин предлагает ввести в формулу Сводеша две поправки, одна из которых отражает замедление процесса замен, связанное с проявлением в списке наиболее стабильной

[6] После исключения из рассмотрения выделяющегося значения λ_1=0,051, (строка 2 табл. 5), найденный при калибровке коэффициент λ=0,16 не изменился.

Рисунок 5. Сравнение исходной $N_{St}(t)$ и калиброванной $N_{StC}(t)$ моделей Старостина: $N_{St}(t) = e^{-0,05 \cdot N_{St} \cdot t^2}$; $N_{StC}(t) = e^{-0,11 \cdot N_{StC} \cdot t^2}$.

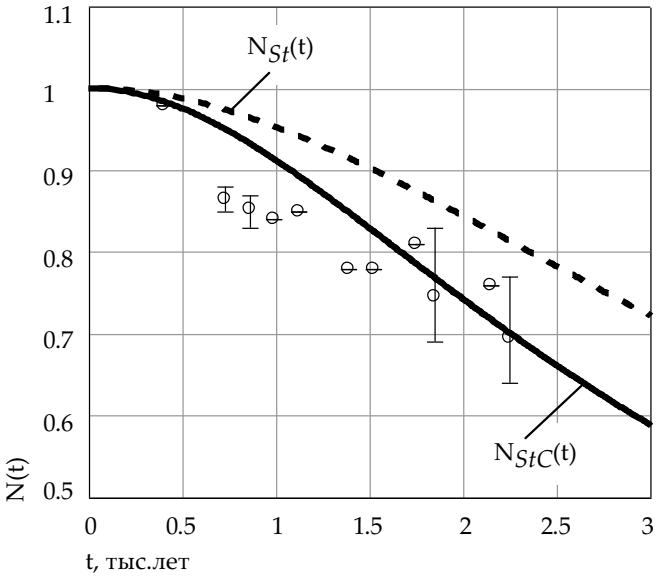
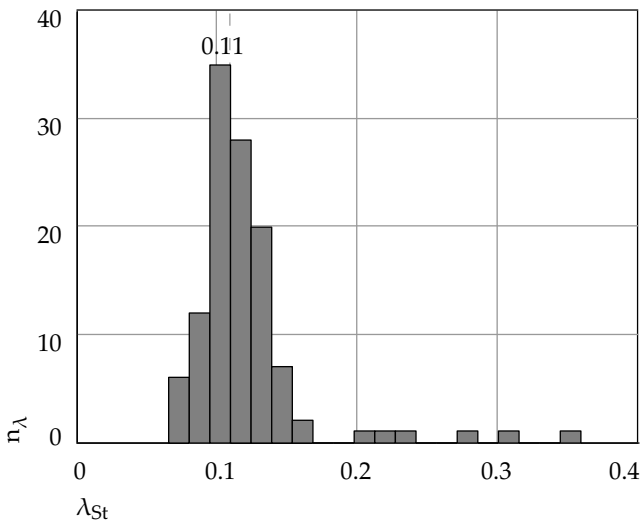


Рисунок 6. Количество пар языков (n_λ) с коэффициентом λ, полученным для калиброванной модели Старостина по исходным данным (табл. 5).



части лексики ($\lambda=\lambda \cdot N(t)$), а вторая — наоборот, его ускорение, обусловленное устареванием сохранившихся значений ($\lambda=\lambda \cdot t$). В результате новая глоттохронологическая модель принимает следующий вид:

$$N_{St}(t) = e^{-\lambda \cdot N_{St} \cdot t^2}.$$

При этом величина константы λ, согласно С. А. Старостину, должна составлять около 0,05 для большинства языков:

$$N_{St}(t) = e^{-0,05 \cdot N_{St} \cdot t^2}.$$

Для калибровки параметров рассмотренной модели по данным романских языков, как и в предыдущем случае, воспользуемся методом наименьших средних квадратов и получим следующие значения коэффициента λ и минимального суммарного отклонения ε:

$$\lambda=0{,}108;$$
$$\varepsilon=0{,}214.$$

Очевидно, что найденный коэффициент λ≈0,11 более чем в два раза отличается от исходного (0,05), что заставляет нас перейти к уточненной модели вида:

$$N_{StC}(t) = e^{-0{,}11 \cdot N_{StC} \cdot t^2}.$$

Как показывает рассмотрение графиков (рис. 5), переход к калиброванной модели позволяет добиться гораздо лучшего совпадения расчетных значений с опорными точками в диапазоне от 1,5 до 2,5 тыс. лет (в отличие от исходной формулы, дающей заметно «заглубленные» датировки практически на любом временном интервале).

Распределение значений λ, полученное по исходным данным для модели Старостина выглядит менее равномерным, чем в случае с формулой Сводеша, однако также достаточно близко к нормальному: математическое ожидание составляет $\overline{\lambda_{St}}$=0,11, среднее квадратическое отклонение σ_λ≈0,02 (рис. 6). При этом наличие шести сильно выделяющихся значений, соответствующих λ>0,2 (точки 3, 4 и 5 на рис. 2), не повлияло на результат калибровки и конечный вид модели[7].

Наряду с описанными выше традиционными глоттохронологическими моделями М. Сводеша и С. А. Старостина рассмотрим также статистический метод, основанный на потоковой интерпретации процесса лексических замен.

1.3. Потоковая глоттохронологическая модель.

В основе потоковой модели лежит представление о том, что процесс замен каждого из значений основного списка является потоком случайных событий, которые происходят с малой интенсивностью и не влияют одно на другое. Таким образом, развитие всего основного списка можно описать как сумму нескольких независимых экспоненциальных потоков с различным, но постоянным темпом распада[8]. Каждая из таких составляющих соответствует группам значений или отдельным значениям в составе основного списка, которые обладают одинаковой или близкой устойчивостью При этом число групп и их коэффициенты стабильности подбираются в зависимости от данных, используемых для калибровки модели. Например, принимая исходное количество составляющих модели равное трем, получаем выражение следующего вида:

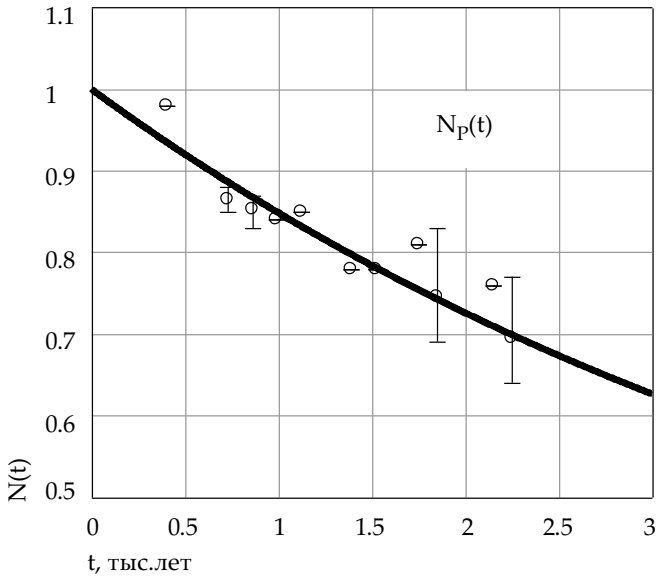
$$N_P(t) = c_1 e^{-\lambda_1 t} + c_2 e^{-\lambda_2 t} + c_3 e^{-\lambda_3 t},$$

где c_1, c_2 и c_3 — доля слов в каждой из составляющих основного списка, а λ_1, λ_2 и λ_3 — соответствующие коэффициенты стабильности каждой из них. При этом сумма всех най-

[7] После исключения всех выделяющихся значений из расчетов величина λ_{St} по-прежнему составила 0,11.

[8] Подробное теоретическое обоснование предлагаемой модели приводится в статье Васильев, Милитарев 2008: 518—523, а практические примеры ее применения в работах Васильев, Старостин 2013, Васильев, Коган 2014. Можно заметить, что данный подход, как и методика С. А. Старостина, подразумевает отказ от 3-го и 4-го постулатов Сводеша, однако использует при этом совершенно другие исходные посылки.

Рисунок 7. Соответствие между полученной потоковой моделью и исходными данными: $N_P(t) = 0{,}2 + 0{,}8e^{-0{,}21 \cdot t}$.



денных компонент $c_1+c_2+c_3$ должна равняться единице (т.е. образовывать полный основной список), а все коэффициенты — иметь неотрицательные значения. Так, например, классическую формулу Сводеша можно представить как частный случай потоковой модели, которая содержит всего одну значимую составляющую $c_1=1$ с коэффициентом $\lambda_1=0{,}16$.

Определим оптимальные параметры потоковой модели по исходным данным (табл. 5 и рис. 2) с помощью метода наименьших средних квадратов. Полученные значения приводятся ниже:

$c_1=0{,}200$; $c_2= 0{,}238$; $c_3=0{,}562$;
$\lambda_1=0{,}000$; $\lambda_2= 0{,}210$; $\lambda_3=0{,}210$;
$\varepsilon=0{,}094$.

Нулевой коэффициент первой составляющей ($\lambda_1=0$), указывает на высокую устойчивость входящих в нее значений (около 20% основного списка), которые сохраняются в языке с течением времени. В то же время равенство коэффициентов $\lambda_2=\lambda_3=0{,}210$ свидетельствует о том, что заданное число слагаемых было избыточным, и модель без потери точности может быть сведена к двум содержательным компонентам: 20% ($c_1=0{,}2$) — сверхстабильная часть списка, 80% ($c_2+c_3=0{,}8$) — экспоненциально убывающая часть значений с коэффициентом $\lambda_{2,3}= 0{,}21$:

$$N_P(t) = 0{,}2 + 0{,}8e^{-0{,}21 \cdot t}.$$

Представленный график (рис. 7) позволяет убедиться в хорошем соответствии потоковой модели фактическим данным о процессе замен в романских языках, что численно подтверждается незначительной суммарной ошибкой ($\varepsilon=0{,}094$), полученной в ходе калибровки.

Завершив рассмотрение основных глоттохронологических методов и моделей, перейдем к сравнению и анализу полученных результатов.

Рисунок 8. Сравнение исходных и калиброванных моделей глоттохронологии, полученных по данным романских языков:

$N_{Sw}(t) = e^{-0,16 \cdot t}$ — исходная/калиброванная модель Сводеша,

$N_{St}(t) = e^{-0,05 \cdot N_{St} \cdot t^2}$ — исходная модель Старостина,

$N_{StC}(t) = e^{-0,11 \cdot N_{StC} \cdot t^2}$ — калиброванная модель Старостина,

$N_P(t) = 0,2 + 0,8\,e^{-0,21 \cdot t}$ — калиброванная потоковая модель.

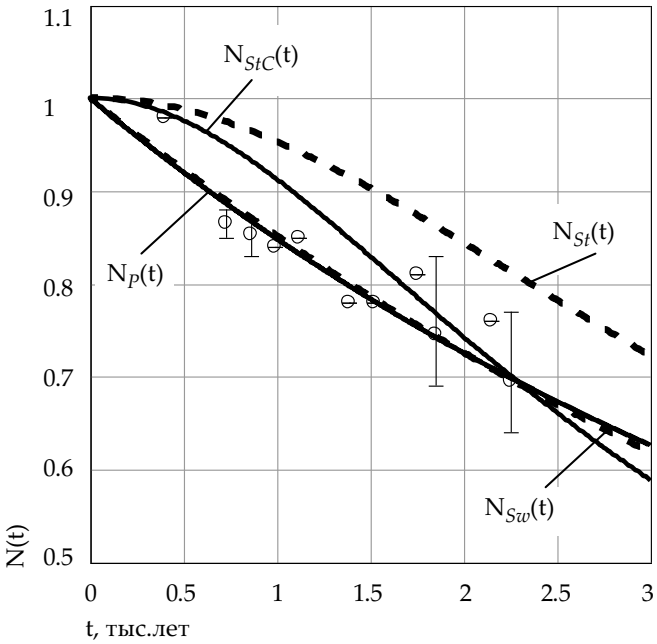


Таблица 6. Сравнение исходных и калиброванных моделей.

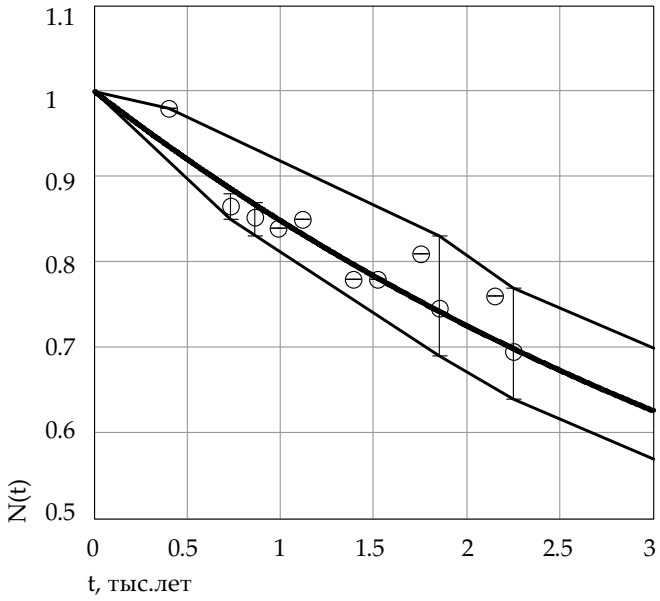
Общий вид и название модели	Вид исходной модели	Вид калиброванной модели
$N(t) = e^{-\lambda \cdot t}$ модель Сводеша	$N_{Sw}(t) = e^{-0,16 \cdot t}$ $\lambda_{Sw} = 0,16$	$N_{SwC}(t) = e^{-0,16 \cdot t}$ $\lambda_{SwC} = 0,16$ ($\varepsilon_{Sw} = 0,094$)
$N_{St}(t) = e^{-\lambda \cdot N_{St} \cdot t^2}$ модель Старостина	$N_{St}(t) = e^{-0,05 \cdot N_{St} \cdot t^2}$ $\lambda_{St} = 0,05$	$N_{StC}(t) = e^{-0,11 \cdot N_{StC} \cdot t^2}$ $\lambda_{StC} = 0,11$ ($\varepsilon_{St} = 0,214$)
$N_P(t) = c_1 e^{-\lambda_1 t} + c_2 e^{-\lambda_2 t} + c_3 e^{-\lambda_3 t}$ потоковая модель	$N_P(t) = 0,2 + 0,8\,e^{-0,21 \cdot t}$ $c_1 = 0,20$; $c_2 = 0,80$; $\lambda_1 = 0,00$; $\lambda_2 = 0,21$; ($\varepsilon_P = 0,094$)	

1.4. Сравнение полученных моделей и их оценка.

Обратимся к сравнительной таблице (табл. 6), которая содержит как исходные, так и калиброванные модели, а также значения соответствующих параметров.

Сопоставление столбцов таблицы показывает, что существенные изменения в ходе калибровки моделей по опорным точкам произошли только в формуле Старостина: полученный коэффициент λ_{St} составил 0,11 при исходном значении 0,05. При этом калиброванная модель Сводеша оказалась неизменной с константой λ_{Sw}=0,16.

Рисунок 9. Фактический разброс значений N(t), используемых в качестве исходных данных для калибрования моделей.



Если мы сравним графики представленных моделей (рис. 8), то обнаружим, что потоковая модель и модель Сводеша (несмотря на принципиальное отличие используемых подходов), дают практически идентичные датировки на всем временном диапазоне с одинаковой суммарной погрешностью $\varepsilon_{Sw}=\varepsilon_P=0{,}094$.

При этом, как следует из рисунка, обе эти модели обеспечивают наилучшее совпадение с опорными точками, в то время как применение модели Старостина приводит к существенным неточностям как с исходным коэффициентом ($\lambda_{St}=0{,}05$), так и после его корректировки в соответствии с исходными данными ($\lambda_{StC}=0{,}11$). И в том, и в другом случае расчетные датировки в интервале 0,5…1,5 тыс. лет оказываются завышенными по отношению к фактическим. Главной причиной этих расхождений является неравномерный «замедленно-ускоренный» характер распада, вызванный введением ускоряющей и замедляющей поправок.

Таким образом, можно сделать вывод, что процесс изменений в базисной лексике одного языка наиболее корректно описывается с помощью экспоненциальной зависимости с постоянным (но не обязательно одинаковым!) темпом замен отдельных значений или частей списка.

Убедившись в принципиальной адекватности математического аппарата глоттохронологии для описания процесса словарных замен, мы можем перейти к численной оценке точности и надежности различных моделей при датировании лексической дивергенции.

1.5. Оценка точности глоттохронологических моделей.

Представляется очевидным, что точность расчетных датировок в первую очередь обусловлена точностью исходных данных, которые используются для идентификации и калибровки параметров моделей. Следовательно, измерение точности следует начинать с оценки фактического разброса значений, объективно присутствующего в исходных данных в силу их статистического характера. Графически данный разброс можно представить в виде кривых, соединяющих крайние значения опорных точек на всем интервале времени: рис. 9.

Рисунок 10 а, б. Иллюстрация разброса фактических долей совпадений и соответствующих датировок по отношению к расчетным значениям N(t) и времени t, полученным по калиброванной модели $N_{Sw}(t) = e^{-0,16 \cdot t}$.

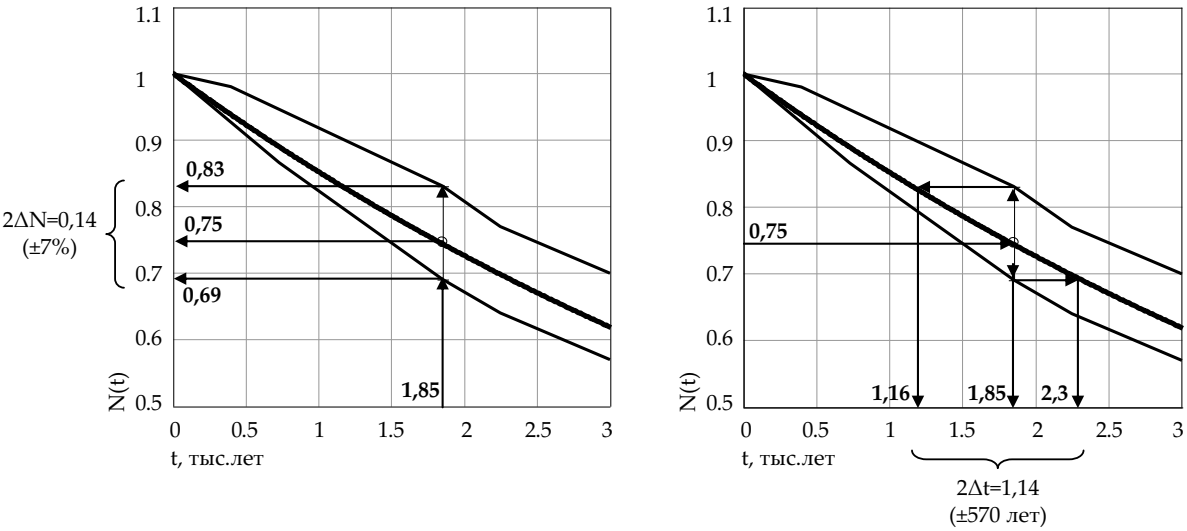
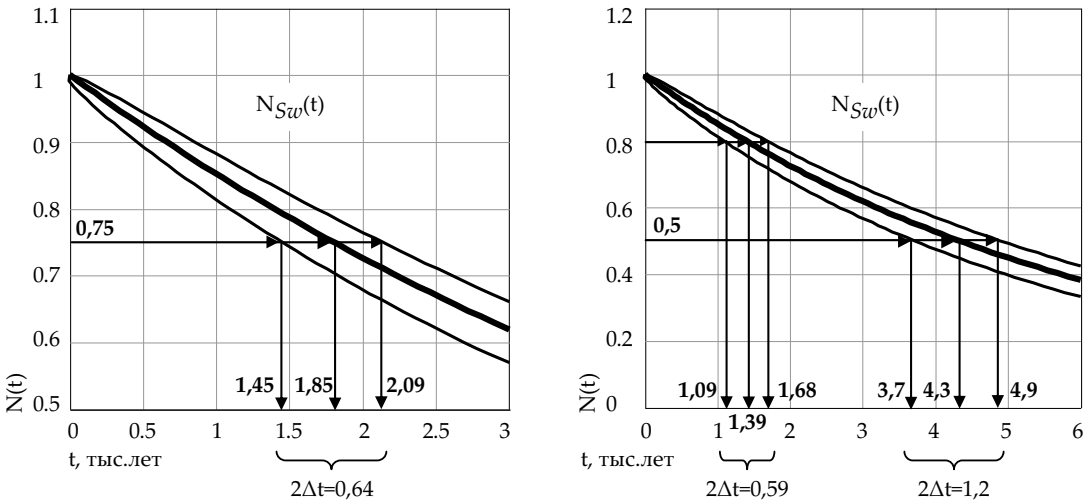


Рисунок 11 а, б. Доверительный интервал, рассчитанный для усредненной модели $N_{Sw}(t) = e^{-0,16 \cdot t}$ с коэффициентом λ=0,16 и заданной вероятностью 0,7.



Используя диаграммы на рис. 10 (а и б), можно численно оценить разброс процентов совпадений для выбранного значения времени или наоборот — интервал неопределенности датировки для известного процента совпадений между сравниваемыми языками. Например, для временного отрезка 1850 лет разброс доли совпадающих значений составит около 14% (т.е. в среднем ±7 слов при использовании 100 словных списков). Аналогичным образом для доли совпадений N(t)=0,75 диапазон временной неопределенности составит 1140 лет (±570 лет).

Изложенный способ оценки точности глоттохронологических моделей подразумевает наличие достаточно большого объема фактических данных о разбросе процентов

Рисунок 12. Доверительный интервал, рассчитанный для усредненной модели $N_{Sw}(t) = e^{-0,16 \cdot t}$ с коэффициентом λ=0,16 и заданной вероятностью 0,95.

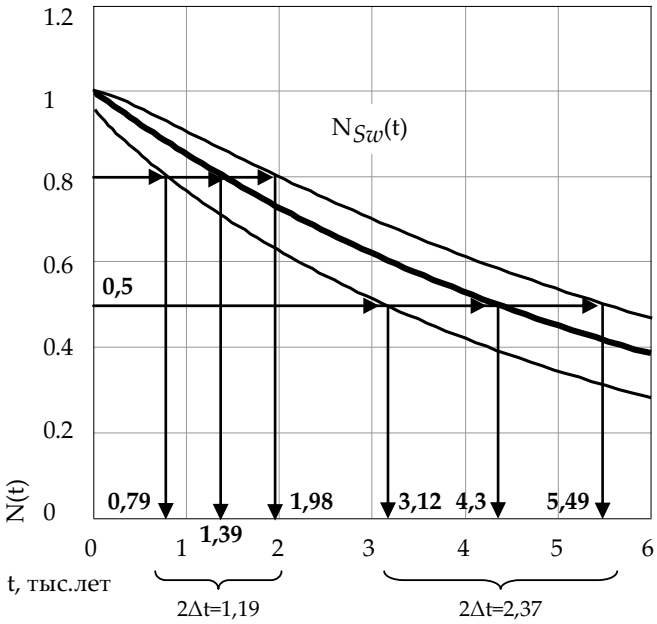


Таблица 7. Зависимость величины доверительного интервала от выбранного времени и вероятности. Приведены усредненные значения. В скобках указан процент величины интервала от значения расчетной датировки.

Расчетная датировка, лет	1000	2000	3000	4000	5000	6000
Величина доверительного интервала для вероятности p = 0,7	±250 (25%)	±360 (18%)	±470 (16%)	±560 (14%)	±650 (13%)	±750 (13%)
Величина доверительного интервала для вероятности p = 0,95	±500 (50%)	±720 (36%)	±940 (32%)	±1120 (28%)	±1300 (26%)	±1500 (25%)

совпадений сравниваемых списков и соответствующих достоверных датировок — т.е. является эмпирическим.

Вместе с тем, на основе описанного выше потокового подхода оценка точности моделей может быть произведена теоретическими методами — благодаря известным статистическим свойствам этого процесса. А именно, если представить общий поток замен в одном списке как сумму потоков замен его значений (каждый из которых является экспоненциальным), то для суммарного потока первых замен по известным формулам можно рассчитать *доверительный интервал* (Вентцель, Овчаров 1969: 235) значений, полученных с использованием модели. Зная величину доверительного интервала, вычисленную для некоторого процента совпадений, мы можем указать диапазон времени, в который с заданной вероятностью[9] укладывается искомая датировка (рис. 11).

Например, при известной доле совпадений N(t)=0,75, пользуясь моделью $N(t) = e^{-0,16 \cdot t}$, получаем расчетную датировку 1850 лет и доверительный интервал 640 лет, из чего сле-

[9] Например, вероятность p=0,7 указывает на то, что в 70 случаях из 100 фактическая датировка будет находиться в пределах нижней и верхней границы рассчитанного доверительного интервала.

дует, что фактическое значение с вероятностью 0,7 может варьироваться в диапазоне от 1450 лет до 2090 лет (рис. 11а). С увеличением временной дистанции ширина доверительного интервала также будет расти. Так, для значения N(t)=0,5 и той же вероятности (p=0,7) его величина достигает уже 1200 лет (рис. 11б). Однако следует отметить, что в процентном отношении величина доверительного интервала постепенно снижается по мере удревнения датировок: например, от 25% (для t=1000 лет) до 13% (для t > 4 тыс. лет), (см. табл. 7, p=0,7). Это обстоятельство хорошо объясняет трудности глоттохронологического сопоставления списков языков с малыми глубинами расхождений.

Численно оценить величину доверительного интервала для разных значений времени и выбранной вероятности можно с помощью табл. 7 и рис.12.

Сопоставив рис. 11 с рис. 12 и значениями из табл. 7, можно убедиться, что фактический разброс значений хорошо совпадает с теоретическим доверительным интервалом, установленным для вероятности p=0,95, что позволяет сделать вывод о статистической адекватности используемых моделей, а также их практической пригодности для получения лингвистических датировок. При этом, однако, следует помнить, что все результаты вычислений будут иметь вероятностный характер, т.е. чем выше желаемая надежность датировки, тем больше величина ее неопределенности. На практике это означает, что при датировании лексических процессов с использованием глоттохронологии следует говорить не о конкретной дате, а о диапазоне дат с известной вероятностью. Например: «нами получена датировка 2000±360 лет с вероятностью 70%» или «2000±720 лет с вероятностью 95%».

Несмотря на то, что исследование процесса лексических замен, происходящих в списке одного языка с течением времени, имеет большое теоретическое значение и составляет основу любых лексикостатистических методов, его применение на практике весьма ограничено. Действительно, случаи, когда необходимо определить временную дистанцию между языком-предком и его потомком, встречаются в компаративистике довольно редко. Гораздо более распространенной задачей является датирование разделения двух современных языков, обладающих предположительным или установленным генетическим родством. Таким образом, с практической точки зрения было бы интересно провести анализ глоттохронологических моделей, описывающих относительную языковую дивергенцию. Этому анализу, как уже говорилось выше, будет посвящена вторая часть настоящей статьи.

Тем не менее, полученные результаты позволяют сформулировать некоторые важные выводы уже по итогам первой части проведенного исследования:

1. Процесс лексических изменений, наблюдаемый в базисной лексике романских языков, наиболее корректно описывается экспоненциальной зависимостью с постоянной скоростью замен отдельных значений или частей списка. Данная зависимость реализована, в частности, в классической модели М. Сводеша, а также, в общем случае, — в потоковой модели.
2. Калибровка параметров глоттохронологических моделей позволяет добиться хорошего соответствия получаемых датировок исходным данным. При этом отдельные фактические значения могут существенно отличаться от расчетных, что свидетельствует, с одной стороны, о ярко выраженной статистической природе и неравномерности лексического процесса, а с другой — о недостаточном количестве опорных точек, используемых при калибровке.
3. Лингвистические датировки, полученные с использованием глоттохронологических методов, имеют *вероятностный характер* и представляют собой не точную величину, а диапазон (доверительный интервал) значений, к которому с известной вероятностью будет принадлежать искомая дата.

4. Величина доверительного интервала может быть определена по фактическим исходным данным, а также рассчитана теоретически — на основе установленных статистических свойств процесса лексических замен. При этом она зависит от измеряемого временного отрезка и желаемой надежности расчетных значений. По мере их увеличения доверительный интервал растет, а точность датировки, соответственно, снижается.
5. Повышение надежности лингвистических датировок, получаемых с использованием глоттохронологии, возможно за счет привлечения дополнительных исходных данных из различных языковых групп и семей, что является важной задачей для будущих лексикостатистических исследований.

Дополнительные материалы доступны на:

- http://jolr.ru/

Файл MS Excel содержит:

- Таблица 8. Проценты совпадений между 110-словными списками романских идиомов.

Литература

Арапов, М. В., М. М. Херц. 1974. *Математические методы в исторической лингвистике*. Москва: Наука.

Васильев, М. Е., А. Ю. Милитарев. 2008. Глоттохронология в сравнительно-историческом языкознании. Модели дивергенции языков. *Orientalia et Classica: Труды Института восточных культур и античности* 19: 509—536.

Васильев, М. Е., А. И. Коган. 2013. К вопросу о восточнодардской языковой общности. *Journal of Language Relationship* 10: 149—177.

Васильев, М. Е., Г. С. Старостин. 2014. Лексикостатистическая классификация нубийских языков: к вопросу о нильско-нубийской языковой общности. *Journal of Language Relationship* 12: 51—72.

Вентцель, Е. С., Л. А. Овчаров. 1969. *Теория вероятностей*. Москва: Наука.

Сводеш, М. 1960а. Лексикостатистическое датирование доисторических этнических контактов. *Новое в лингвистике* 1: 23—52.

Сводеш М. 1960б. К вопросу о повышении точности в лексикостатическом датировании. Новое в лингвистике 1: 53—87.

References

Arapov, M. V., M. M. Herz. 1974. *Matematicheskiye metody v istoricheskoy lingvistike*. Moskva: Nauka.

Bergsland, K., H. Vogt. 1962. On the validity of glottochronology. *Current anthropology* 3(2): 115—153.

Kroeber, A. L. 1958. Romance history and glottochronology. *Language* 34(4): 454—457.

Rea, J. A. 1958. Concerning the Validity of Lexicostatistics. *International Journal of American Linguistics* 24(2): 145—150.

Starostin, S. 2000. Comparative-historical linguistics and lexicostatistics. In Colin Renfrew et al. (eds.). *Time Depth in Historical Linguistics*. Cambridge: McDonald Institute for Archaeological Research. Vol. 1: 233—259.

Swadesh, M. 1952. Lexico-statistic dating of prehistoric ethnic contacts. *Proceedings of the American philosophical society* 96(4): 452—463.

Swadesh, M. 1960a. Leksikostatisticheskoye datirovaniye doistoricheskikh etnicheskikh kontaktov. *Novoye v lingvistike* 1: 23—52.

Swadesh, M. 1960b. K voprosu o povyshenii tochnosti v leksikostatisticheskom datirovanii. *Novoye v lingvistike* 1: 53—87.

Vasilyev, M. E., A. I. Kogan. 2013. K voprosu o vostochnodardskoy yazykovoy obshchnosti. *Journal of Language Relationship* 10: 149—177.

Vasilyev, M. E., A. Yu. Militaryov. 2008. Glottokhronologiya v sravnitel'no-istoricheskom yazykoznanii. Modeli divergentsii yazykov. *Orientalia et Classica: Trudy Instituta vostochnykh kultur i antichnosti* 19: 509—536.

Vasilyev, M. E., G. S. Starostin. 2014. Leksikostatisticheskaya klassifikatsiya nubiyskikh yazykov: k voprosu o nil'sko-nubiyskoy yazykovoy obshchnosti. *Journal of Language Relationship* 12: 51—72.

Venzel, E. S., L. A. Ovcharov. 1969. *Teoriya veroyatnostey*. Moskva: Nauka.

Mikhail E. Vasilyev, Mikhail N. Saenko. How accurate glottochronology can be? Dating the lexical replacement process in the Romance languages.

In this paper we discuss the accuracy of glottochronology, a lexicostatistical method used in the dating of linguistic divergence. Our study provides a detailed analysis of the process of lexical replacement in the basic lexicon of one language over the course of time. To measure replacement rates and determine other statistic features of lexical change we use 110-item wordlists, compiled over the past two years for 54 modern and several historically attested Romance languages. Pairwise comparison of modern wordlists with those of Archaic Latin, Late Classical Latin, Old French, and Old Italian allows to obtain several control points suitable for calibration of glottochronological equations. To estimate the time distance between the compared idioms, three different methods have been applied: the classic formula of M. Swadesh, the modified glottochronology of S. Starostin and a recently proposed approach based on simulation of lexical changes of every meaning on the Swadesh list as stationary Poisson processes. Further analysis resulted in several important conclusions concerning the following questions: (a) what are the main characteristics of lexical divergence in one language; (b) which of the existing models maps these characteristics more efficiently; (c) how precise and reliable glottochronological dating can be in general. We plan to follow this research by another study in which the process of relative divergence between two or more languages with the same ancestor will be considered.

Keywords: Romance languages, lexicostatistics, glottochronology, Swadesh wordlist.