Модифицированная глоттохронология

Метод Старостина

Distanz zwischen slawischen Sprachen, SoSe 2010
16. Juni 2010

Lehner Olga

Основные постулаты глоттохронологии Мориса Сводеша [Арапов-Херц 1974]

1. В словаре каждого языка можно выделить специальный фрагмент, который называется основной или стабильной частью.
2. Можно указать список значений, которые в любом языке обязательно выражаются словами из основной части. Эти слова образуют основной список (ОС). Через N0 обозначим число слов в ОС.
3. Доля p слов из ОС, которые сохранятся (не будут заменены другими словами) на протяжении интервала времени t, постоянна (т.е. зависит только от величины выбранного промежутка, но не от того, как он выбран или какие слова какого языка рассматриваются).
4. Все слова, составляющие ОС, имеют одинаковые шансы сохраниться (соответственно, не сохраниться, "распасться") на протяжении этого интервала времени.
5. Вероятность для слова из ОС праязыка сохраниться в ОС одного языка-потомка не зависит от его вероятности сохраниться в аналогичном списке другого языка-потомка.

Время расхождения

$$t = \frac{\ln(c)}{-L}$$

где
t — данный период времени от одной стадии языка к другой,
c — процент сохранившихся к концу этого периода единиц из списка,
L — скорость замены для этого списка слов. (L=1-r)

Сводеш (1960)
получил эмпирическое значение приблизительно 0.14 для L (означающее, что скорость замены составляет около 14 слов из 100-словного спика в тысячелетие).
N(t) доля слов исходного ОС, сохранившихся к моменту t:

$$N(t) = N_0 e^{-Lt}$$

Противоречия

- Bergsland and Vogt (1962): для скандинавских языков скорость распадения лексики за последнюю тысячу лет в исландском языке равнялась всего ~0,04, а в литературном норвежском (риксмола) - ~0,2. Тогда получаются совершенно нелепые результаты: для исландского языка - около 100-150 лет, а для норвежского - 1400 лет самостоятельного развития, хотя из исторических данных известно, что оба языка развивались из одного источника и существовали независимо около 1000 лет.
- Языковые изменения происходят из-за социальных и исторических событий, которые, разумеется, являются непредвиденными, и, таким образом, не поддаются строгому анализу.

Метод Старостина

Старостин, Сергей Анатольевич (1953-2005)

- Исключение заимствованных слов Систематически заимствуемые слова, которые заимствуются одним языком из другого, являются нарушающим фактором и должны быть исключены из вычислений; „Ускоренное“ развитие риксмола обьясняется заимствованиями, его стословный список включает 11 датских, 3 шведских и 2 немецких заимствования. Исключение этих элементов из расчетов снижает значение до ожидаемой скорости в 5-6 «родных» замен за тысячелетие
- отказ от третьего постулата Скорость изменения L, в действительности, не постоянна, но зависит от периода времени t, в течение которого слово существует в языке (то есть вероятность замены лексемы X лексемой Y возрастает прямо пропорционально прошедшему времени — так называемому «старению слов» . Формула

$$N(t) = N_0 e^{-Lt^2}$$

- несколько улучшает датировку для близких эпох, но по мере углубления в древность дает еще худшие результаты чем Сводешевская глоттохронология.

Метод Старостина

нарушение 4-го постулата Отдельные единицы в 100-словном списке имеют другой уровень стабильности (например, для слова «я» обычно вероятность замены намного ниже, чем для слова «желтый» и т. д.)

$$N(t) = N_0 e^{-L * N(t) * t^2}$$

Формула отражает „противоречивый“ характер процесса распада лексики в ОС: показатель степени при t отражает ускорение распада по мере „старения“ слов, а коэффициент N(t) в показателе степени отражает, напротив, замедление скорости распада по мере выпадения из исходного ОС менее устойчивых слов и сохранение более устойчивых.

Для практического использования (в наиболее частом случае), когда N0=1 (100 слов) и более традиционном обозначении N(t) имеем:

$$t = \sqrt{\frac{\ln c}{-Lc}}$$

Генеалогическое дерево языков

- В.В.Кромер (2003) на основании матрицы коэффициентов совпадений между основными списками словарей строил дендрограммы (генеалогическое дерево языков)
- pdf Глоттохронология и проблемы праязыквой реконструкции//Когнитивное моделирование в лингвистике: сб.докл. Вып. 8. М.: МИСиС, 2003. – С. 238-252

Расстояние Левенштейна

Расстояние Левенштейна (также редакционное расстояние или дистанция редактирования) между двумя строками в теории информации и компьютерной лингвистике — это минимальное количество операций вставки одного символа, удаления одного символа и замены одного символа на другой, необходимых для превращения одной строки в другую.

Расстояние Левенштейна и его обобщения активно применяется

- для исправления ошибок в слове (в поисковых системах, базах данных, при вводе текста, при автоматическом распознавании отсканированого текста или речи).
- для сравнения текстовых файлов утилитой diff и ей подобными. Здесь роль «символов» играют строки, а роль «строк» — файлы.
- в биоинформатике для сравнения генов, хромосом и белков.

см. Википедия

Редакционное предписание

Редакционным предписанием называется последовательность действий, необходимых для получения из первой строки второй кратчайшим образом.

Обычно действия обозначаются так: D (англ. delete) — удалить, I (англ. insert) — вставить, R (replace) — заменить, M (match) — совпадение.

Например, для 2-х строк «CONNECT» и «CONEHEAD» можно построить следующую таблицу преобразований:

M	M	M	R	R	R	R	I
C	O	N	N	E	C	T	
C	O	N	E	H	E	A	D

Разные цены операций

Цены операций могут зависеть от вида операции (вставка, удаление, замена) и/или от участвующих в ней символов, отражая разную вероятность разных ошибок при вводе текста и т. д. В общем случае:

- $w(a, b)$ — цена замены символа a на символ b
- $w(\varepsilon, b)$ — цена вставки символа b
- $w(a, \varepsilon)$ — цена удаления символа a

Необходимо найти последовательность замен, минимизирующую суммарную цену. Расстояние Левенштейна является частным случаем этой задачи при

- $w(a, a) = 0$
- $w(a, b) = 1$ при $a \neq b$
- $w(\varepsilon, b) = 1$
- $w(a, \varepsilon) = 1$

Транспозиция

Если к списку разрешённых операций добавить транспозицию (два соседних символа меняются местами), получается расстояние Дамерау — Левенштейна. Для неё также существует алгоритм, требующий O(MN) операций.

Дамерау показал, что 80 % ошибок при наборе текста человеком являются транспозициями.

Литература

- **Арапов-Херц**: М.В.Арапов, М.М.Херц. Математические методы в исторической лингвистике. М.:, 1974.
- **Bergsland, Knut; & Vogt, Hans**. (1962). On the validity of glottochronology. *Current Anthropology* , *3* , 115-153.
- **Старостин С.А.** Сравнительно-историческое языкознание и лексикостатистика// Лингвистическая реконструкция и древнейшая история Востока (Материалы к дискуссиям международной конференции). Т.1. М.:Наука, 1989. С. 3-39.
- **В.В. Кромер.** Глоттохронология и проблемы праязыквой реконструкции//Когнитивное моделирование в лингвистике: сб.докл. Вып. 8. М.: МИСиС, 2003. – С. 238-252
- **В. И. Левенштейн**. Двоичные коды с исправлением выпадений, вставок и замещений символов. Доклады Академий Наук СССР, 1965. 163.4:845-848.