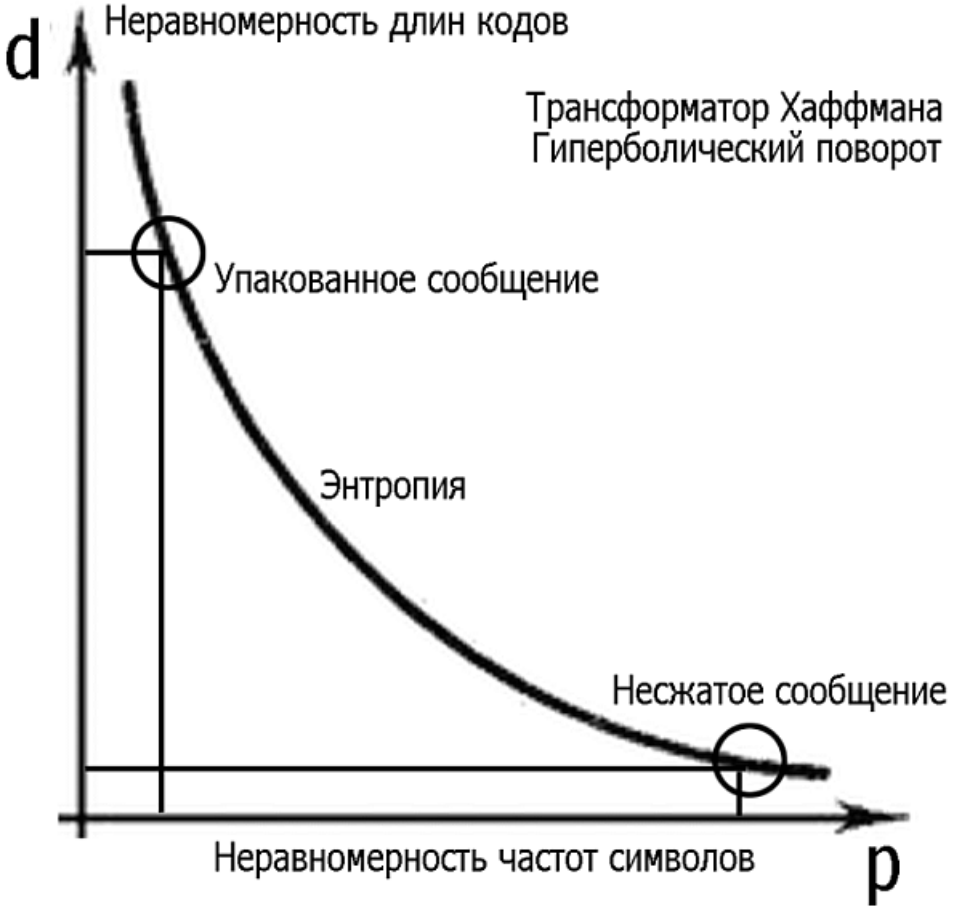
# Диакоптика трансформаторов текста




Все методы трансформации текста естественно распадаются на две группы: диссипативные, связанные с потерей информации (аннотирование, реферирование, компиляция, сжатие данных с потерями - *lossy data compression*) и консервативные, без потерь информации (перевод с одного языка на другой, шифрование, сжатие данных без потерь - *lossless data compression*) .

Из консервативных методов именно техники *lossless data compression* являются наиболее удобным объектом анализа, своего рода дрозофилой лингвистики.

Успех теории относительности Эйнштейна породил интерес к ее математическому аппарату - тензорному анализу. В 30-х годах прошлого века американский инженер-электрик Габриэль Крон опубликовал серию статей о разработанном им методе, позднее названном "Диакоптика", систематически использующим аппарат тензорного исчисления для дискретных систем.

С точки зрения Крона, полная система может быть фрактально разделена на подсистемы меньшего размера и связи между ними. Поведение подсистем может быть изучено независимо и затем они могут быть вновь объединены в полную систему с учетом своих связей. Для облегчения анализа, подсистемы могут быть подвергнуты различным ортогональным преобразованиям, сохраняющим их

энергетическую эквивалентность ("инвариант мощности Крона").

Техники сжатия данных без потерь (*lossless data compression*) сводятся к отображению одной системы кодов в другую, сохраняющие в качестве инварианта "количество информации" по Шеннону. При этом в качестве "сжатого" выбирается преобразование, отвечающее минимальному размеру текста (или, двойственно, его максимальной энтропии).

В терминологии Крона, это означает, что два контрвариантных тензора - кратностей символов сообщения **Q** и битовых размеров кодов символов **W,** связаны между собой через метрический тензор (импеданса) **z**.

**W = z * Q**

Прямое ("упаковка") и обратное ("распаковка") преобразования связаны между собой отношением инверсии.

Точный вид использованного алгоритма не имеет значения, с формальной точки зрения (исключая детали реализации), все они приводят к тому же самому результату.

Всего возможны ровно четыре комбинации различных техник:

- замена равномерных кодов равномерными (BWT, ST)
- замена равномерных кодов неравномерными (Huffman)
- замена неравномерных кодов равномерными (LZ)
- замена неравномерных кодов неравномерными

Замена равномерных кодов равномерными, очевидно, не дает выигрыша в сжатии и используется в качестве "препроцессинга" для других алгоритмов.

Замена равномерных кодов неравномерными - старейшая техника сжатия, впервые использованная Самуэлем Морзе и, позднее, систематически развитая Шенноном-Фано и Хаффманом.

Симметричная к ней замена неравномерных кодов равномерными предложена относительно недавно, и систематически используется в так называемых "словарных" алгоритмах (семейство LZ).

Замена неравномерных кодов неравномерными технически более сложна и, обычно, заменяется каскадной комбинацией предыдущих техник.

Заметим, что формальная обратимость не влечет термодинамической обратимости. Как было сформулировано ранее, текст является квазиупругой (гистерезисной) системой. Иными словами, двойное преобразование требует двойной работы. Надеждам Ландауэра на "консервативные вычисления", очевидно, не суждено сбыться: если дважды перевернуть песочные часы, время не вернется назад.

Существование инварианта трансформации ("количества информации" по Шеннону), означает, что битовый размер текста и его энтропия связаны между собой отношением инверсии. Иными словами, сжатие данных является гиперболическим поворотом в плоскости WQ ("лоренц-сжатие").

Разумеется, шенноновское "количество информации" никакого отношения к понятию "информация" не имеет. Математически, это инвариант формальной системы (сообщения), по физическому смыслу - "энергия деформации сообщения", по Крону - "инвариант мощности", сохраняемый при ортогональном преобразовании.