Размышления. Энтропийные трансформаторы текста

Сам термин "*сжатие*", разумеется, неверен, но устоялся и общепринят. По крайней мере, визуально, это выглядит как сжатие, скажем, при замене регулярного шрифта петитом. И если с помощью подходящего кодирования при электрической связи удается обойтись передачей меньшего числа импульсов тока, то выигрывается и время и деньги.

Теория информации не рассматривает энергетические аспекты взаимодействия систем, а только изменения их состояния. Иными словами, в информационной модели, системы никогда не обмениваются энергией, а все изменения происходят только за счет работы сторонних сил.

Тем не менее, ничто не препятствует и "обычному" (механическому) подходу, с приписыванием деформации "формальной" энергии.

Можно сказать, что "передача сообщения" - это принудительная синхронизация двух функционально идентичных систем с пренебрежимо малым обменом энергией.

Прямое и обратное кодирование (компрессию-декомпрессию) можно рассматривать как упругую (обратимую) деформацию сообщений.

К сожалению, ни термин "*сжатие*", ни термин "*кодирование*" не отражают специфики преобразования данных. Возможно, более правильным было бы использование термина "компиляция" - для обозначения, в общем случае, двухстадийного процесса: лексический анализ + инверсия кодового дерева (энтропийное кодирование).

В тривиальных схемах, лексический разбор может сводиться к "нарезке" входного потока на токены фиксированной битовой ширины (обычно, 4, 8 или 16 бит), а инверсия кодового дерева заменяется просто мэппингом токенов в индексы, без учета их кратности.

Важно, что "сжимаем" только неоднородный текст (с различной кратностью лексем). Полностью однородный текст (часто, но неверно, понимаемый как "случайный") несжимаем. Иными словами, для произвольного алфавита, полная алфавитная последовательность ABC... или любая ее пермутация являются несжимаемыми.