Използване на КА за кодиране на морфологични речници

(Това изложение следва примерите в глава 3 на дисертацията на Стоян Михов, вж. http://lml.bas.bg/~stoyan/diser.ps.gz)

Ако поддържаме морфологичния си речник като списък на основни форми (лексеми) и асоциирани към тях флективни класове, ще ни се налага да разпознаваме словоформите чрез процедури за анализ и синтез всеки път, когато искаме да анализираме или генерираме текст. Вместо това приемаме простата идея да генерираме веднъж речник от всички възможни словоформи и да поддържаме направо него, което не е проблем при сегашното пространство оперативна и дискова памет, и през последните няколко години се счита за стандарт за представяне на речници. Да видим чрез примери как речникът може да се превърне в КА. Да разгледаме следния примерен речник от словоформи, зададен в азбучен ред:

ваза, ваза. N:fs
вазата, ваза. N:fsd
вази, ваза. N:fp
вазите, ваза. N:fpd
вода, вода. N:fs
водата, вода. N:fsd
води, вода. N:fp, водя. V+t+IPR:P3s:A2s:A3s:Z2s
водите, вода. N:fpd, водя. V+t+IPR:P2p
маса, маса. N:fs
масата, маса. N:fsd
маси, маса. N:fp
масите, маса. N:fpd

В началото на реда стои словоформата, следва разделител ",", основната форма и разделител ".". После са дадени граматическите характеристики на формите. С главни латински букви са кодирани граматически категории, а с малки букви и цифри - техни стойности. При наличие на повече от една стойност, всички те са изброени с разделител ":". В този случай: **N** означава съществително, **f** женски род, **s** ед. ч-ло, **p** множ. ч-ло, **d** членувано, **V+t+IMP** - преходен глагол несвършен вид, **P** сег. време, **2s** и **3s** съотв. 2-ро и 3-то лице ед.ч-ло, **A** е мин. време и **Z** - заповедно наклонение. Това е речник в т.нар. DELAF-формат, широко използван след 1993 в известната система INTEX (в момента една по-съвременна версия е наречена GlossaNet и е достъпна на http://glossa.ladl.jussieu.fr/info.html)

Забелязваме, че дори в този прост речник "N:fp" се повтаря три пъти и решаваме да извадим граматическите характеристики в отделен списък.

Нека с **Xu** означим преобразуването: дадена словоформа се получава от основната, като се изтрият **X** букви от края и към останалото се долепи **u**. Речникът добива вида:

ваза, 0. N:fs
вазата, 2. N:fsd
вази, 1а. N:fp
вазите, 3а. N:fpd
вода, 0. N:fs
водата, 2. N:fsd
води, 1а. N:fp, 1я. V+t+IPR:P3s:A2s:A3s:Z2s
водите, 3а. N:fpd, 3я. V+t+IPR:P2p
маса, 0. N:fs
масата, 2. N:fsd
маси, 1а. N:fp
масите, 3а. N:fpd

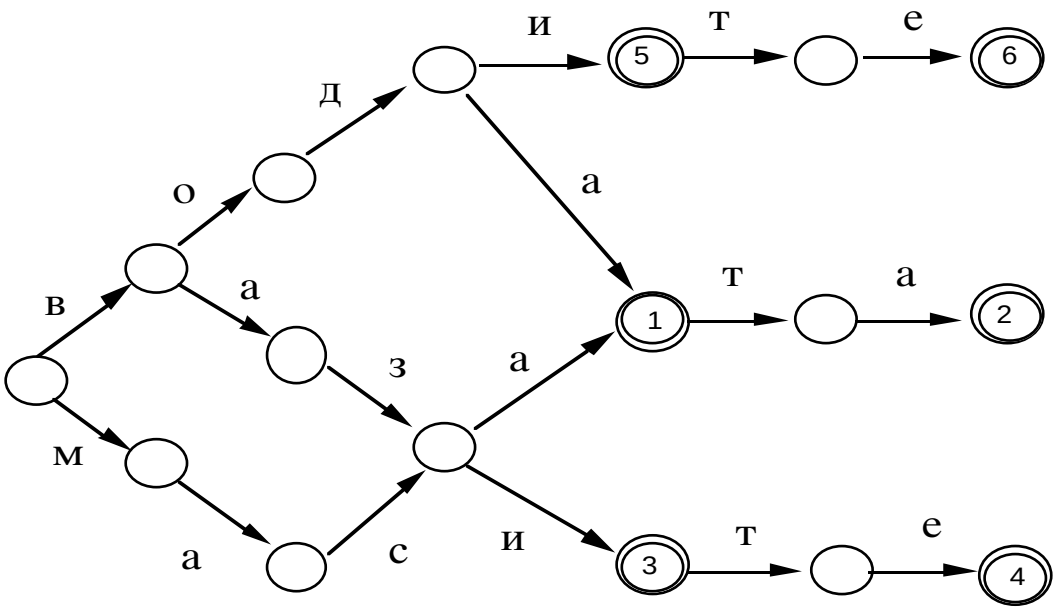
Отделяме характеристиките за получаване на формата и граматическата й информация в отделен списък с цел опростяване на етикетите в речника:

1. 0. N:fs
2. 2. N:fsd
3. 1а. N:fp
4. 3а. N:fpd
5. 1а. N:fp, 1я. V+t+IPR:P3s:A2s:A3s:Z2s
6. 3а. N:fpd, 3я. V+t+IPR:P2p

ваза, 1
вазата, 2
вази, 3
вазите, 4
вода, 1
водата, 2
води, 5
водите, 6
маса, 1
масата, 2
маси, 3
масите, 4

Получени са “класове” думи, като за всеки клас се знае как формата се получава от основната и каква е граматическата й характеристика. В този вид речникът е удобен за представяне чрез ацикличен краен автомат, както

е показано по-долу. На всяка дъга (преход) съответства една буква. Дума с дължина **n** се разпознава за **n** стъпки в автомата. Забележете, че на крайно състояние 1 съответства класа 1 (0. N:fs), което показва, че автоматът е построен по специален начин (по-нататък ще видим как). Долният автомат е минималният с тези свойства. Това е краен автомат с етикети на крайните състояния, моделиращ морфологичен речник.



Количествени характеристики на автомат, моделиращ морфологичен речник на българския език с 60 000 основни форми:

- словоформи 893 313
- състояния на автомата 47 536
- преходи в автомата 110 105
- класове (етикети на крайните състояния) 6244

Обем на речника представен по този начин: лекция 3.