Конечные автоматы

Мати Рейнович Пентус, Алексей Андреевич Сорокин

МГУ им. М. В. Ломоносова
весенний семестр 2022/2023 учебного года
Межфакультетский курс
“Введение в компьютерную лингвистику”

Определение конечного автомата

Пусть Σ — конечный алфавит.

Определение конечного автомата

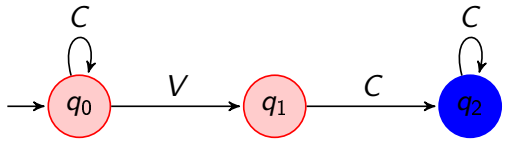
Конечный автомат: кортеж $M = \langle Q, \Sigma, \Delta, q_0, F \rangle$, где

- Q — конечное множество состояний
- $\Delta \subseteq Q \times (\Sigma \cup \{\varepsilon\}) \times Q$ — конечное множество переходов
- $q_0 \in Q$ — стартовое состояние
- $F \subseteq Q$ — завершающие состояния.

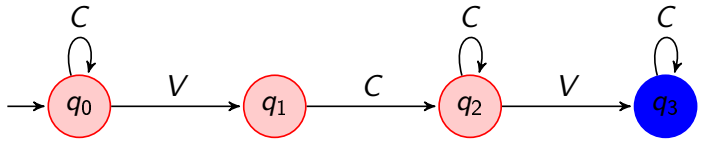
Неформально, конечный автомат — граф с выделенными стартовой и заверщающими вершинами, рёбра которого помечены символами алфавита или пустым словом.
$L(M)$ — метки путей из начального состояния в завершающие.
Язык автоматный — задаётся некоторым конечным автоматом.

Примеры конечных автоматов

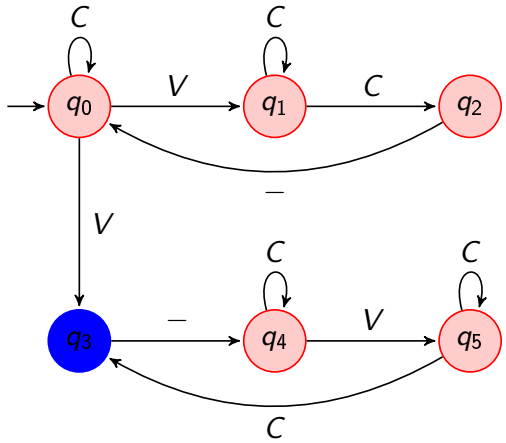
- Закрытый слог



- Слово с 2 гласными, разделёнными хотя бы одним согласным:

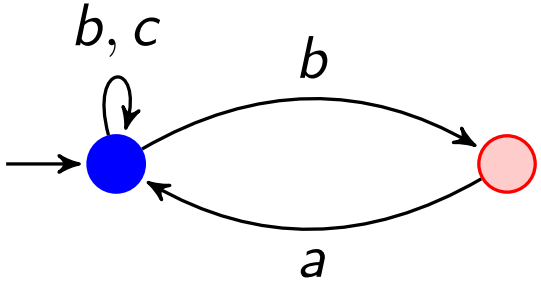


- Слогоделение ровно с одним открытым слогом:

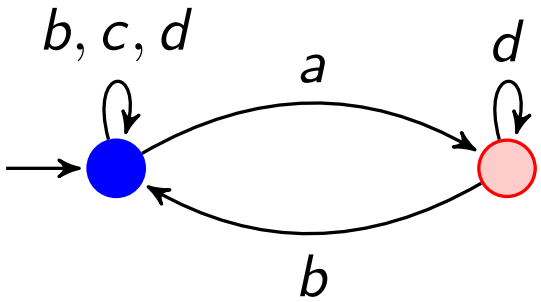


Конечные автоматы: примеры

- Каждой a непосредственно предшествует b, алфавит a, b, c.

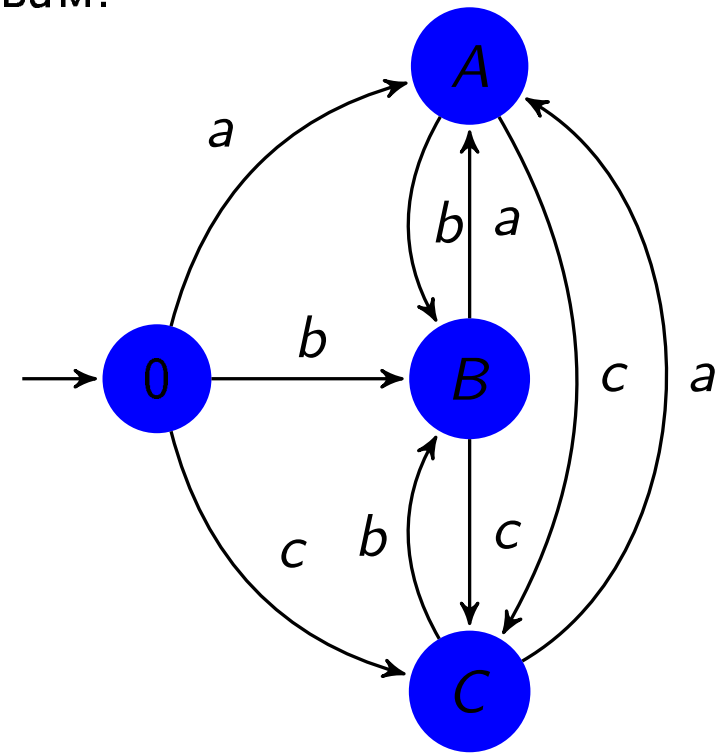


- Справа от каждой a есть парная ей b, между парными буквами нет a, c, алфавит a, b, c, d.



Конечные автоматы: примеры

Нет повторяющихся букв, алфавит a, b, c. Состояния соответствуют буквам:



Автоматы с однобуквенными переходами

Теорема

Каждый автоматный язык распознаётся автоматом с однобуквенными переходами.

Схема доказательства

- Сделать завершающими все состояния, из которых достижимо по ε (возможно, за несколько шагов) другое завершающее.
- Добавить все рёбра вида $\langle q_1, a\rangle \to q_2$,для которых существуют состояние q_3, такое что есть ребро $(\langle q_3, a\rangle \to q_2) \in \Delta$ и ε-путь из q_1 в q_3.
- Удалить ε-рёбра.

Детерминированные конечные автоматы

Определение

Автомат с однобуквенными переходами — детерминированный, если ни из какого состояния не выходит двух рёбер, помеченных одинаковыми буквами.

Теорема

Каждый автоматный язык распознаётся детерминированным автоматом.

Схема доказательства

- Новые состояния — множества старых состояний.
- Ребро, помеченное a, ведёт из Q_1 в Q_2, если Q_2 содержит в точности состояния,достижимые из Q_1 по a.
- Стартовое множество состояний $Q_0 = \{q_0\}$.
- Завершающие состояния: множества, содержащие хотя бы одно завершающее.

Теорема Клини

Теорема

Классы автоматных и регулярных языков совпадают.

Схема доказательства

- Нужно строить по любому конечному автомату эквивалентное регулярное выражение и наоборот: по выражению — автомат.
- Автомат $\to$ выражение: сложно.
- Выражение $\to$ автомат: индукция по построению:
- Регулярные языки строятся из базовых с помощью операций.
- Базовые регулярные языки (буквы и пустое слово) — автоматные.
- Надо доказать, что операции сохраняют автоматность.

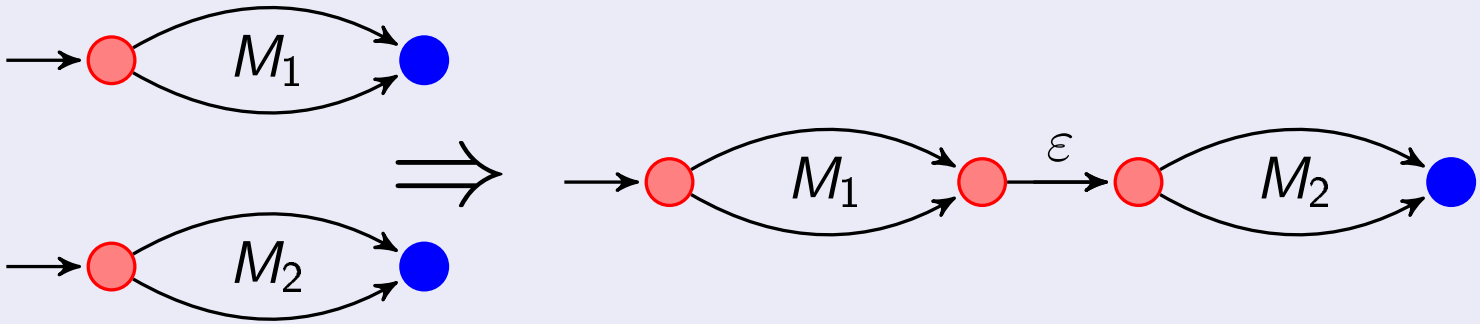
Теорема Клини

Теорема

Классы автоматных и регулярных языков совпадают.

Схема доказательства

Конкатенация: $L_1 = L(M_1), L_2 = L(M_2) \to L_1 \cdot L_2 = L(M)$



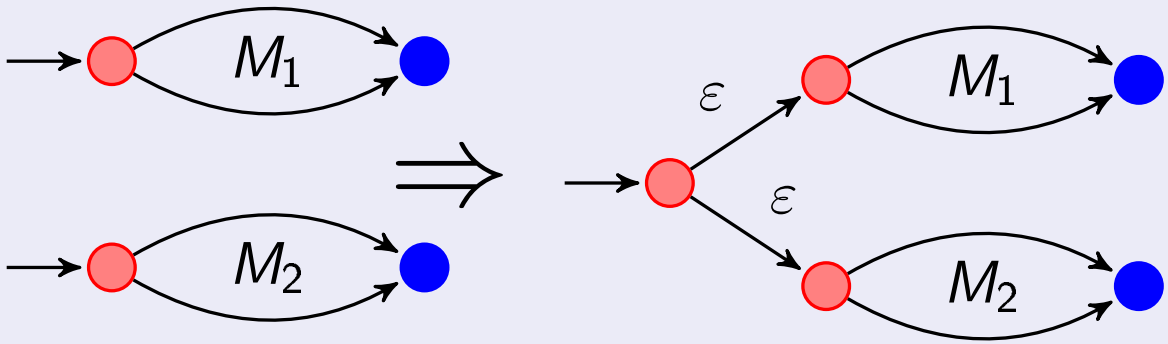
Теорема Клини

Теорема

Классы автоматных и регулярных языков совпадают.

Схема доказательства

Объединение: $L_1 = L(M_1), L_2 = L(M_2) \to L_1 \cup L_2 = L(M)$



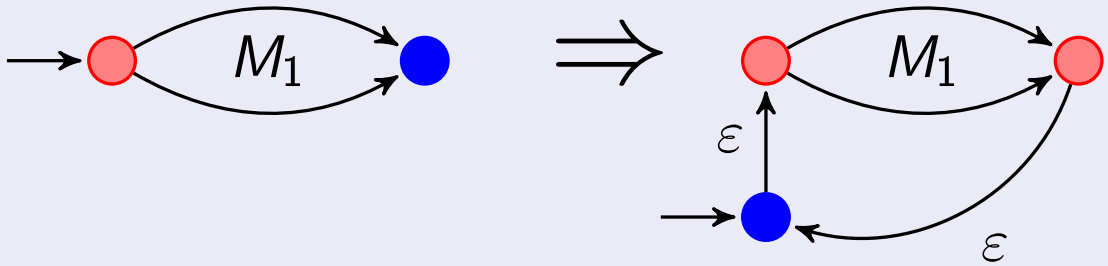
Теорема Клини

Теорема

Классы автоматных и регулярных языков совпадают.

Схема доказательства

Итерация: $L_1 = L(M_1), L_1^* = L(M)$



Дополнение автоматного языка

Теорема

Класс автоматных языков замкнут относительно дополнения.

Идея доказательства

- Рассмотрим детерминированный автомат для языка L.
- Пополним его новым “стоковым” состоянием q'.
- Если из состояния q_1 нет ребра по букве a, добавим ребро $\langle q_1, a\rangle \to q'$.
- Добавим рёбра $\langle q', a\rangle \to a$ для всех a.
- Теперь для всех $q_1 \in Q, a \in \Sigma$ есть ребро вида $\langle q_1, a\rangle \to q_2$.
- То есть каждое слово w приводит из q_0 ровно в 1 состояние.
- Инвертировав завершающие и незавершающие состояния, получим автомат для дополнения.

Пересечение автоматных языков

Теорема

Класс автоматных языков замкнут относительно пересечения.

Идея доказательства

- Короткий вариант: $L_1 \cap L_2 = \overline{\overline{L_1} \cup \overline{L_2}}$.
- Длинный (но эффективный): рассмотрим полные детерминированные автоматы M_1, M_2 для языков L_1, L_2.
- Пусть Q_1, Q_2 их множества состояний, q_{01}, q_{02} — стартовые, а F_1, F_2 — множества завершающих.
- Состояния нового автомата — пары старых состояний $\langle q_1, q_2 \rangle$, $q_1 \in Q_1, q_2 \in Q_2$.
- Новое стартовое — пара старых стартовых $\langle q_{01}, q_{02} \rangle$.
- Автомат симулирует M_1 по первой координате и M_2 — по второй.
- Завершающие состояния — пары завершающих (автомат обязан принимать слово по обеим координатам).

Рекурсивное построение автоматов

- Автоматные языки замкнуты относительно множества операций.
- Это замыкание эффективно: соответствующие автоматы строятся алгоритмически.
- Вывод: можно строить регулярные выражения из автоматов (но с большим числом операций).
- Множественное число в английском:

$$(L_{sib} \cdot es) \cup (((\overline{L_{sib}} \cap L_C) \cup L_{Vy} \cup L_V) \cdot s),$$

где

  - L_{sib} — слова, кончающиеся на шипящий.
  - L_C — слова, кончающиеся на согласный.
  - L_{Vy} — слова кончающиеся на гласный+y.
  - L_V — слова, кончающиеся на гласный (не y).

- $L_{sib}, L_C, L_{Vy}, L_V$ — автоматные, финальный автомат строится рекурсивно.

Конечные преобразователи

Неформально, конечный преобразователь — это автомат с добавленными на рёбра выходными символами.

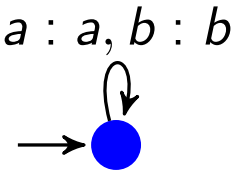
Пусть Σ, Γ — конечные алфавиты.

Определение конечного преобразователя

Конечный преобразователь: кортеж $M = \langle Q, \Sigma, \Gamma, \Delta, q_0, F \rangle$, где

- Q — конечное множество состояний
- $\Delta \subseteq Q \times (\Sigma \cup \{\varepsilon\}) \times (\Gamma \cup \{\varepsilon\}) \times Q$ — конечное множество переходов
- $q_0 \in Q$ — стартовое состояние
- $F \subseteq Q$ — завершающие состояния.

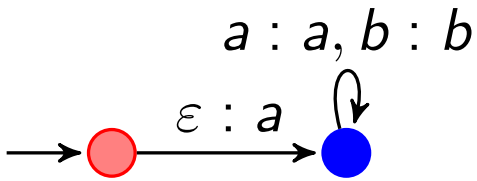
Простейший преобразователь — тождественный (алфавит a, b):



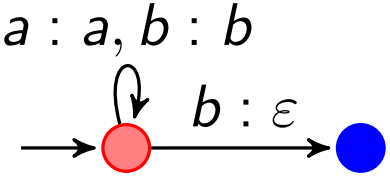
Конечные автоматы можно понимать как преобразователи, возвращающие свой вход (и проверяющие, что вход принадлежит нужному множеству).

Конечные преобразователи: примеры

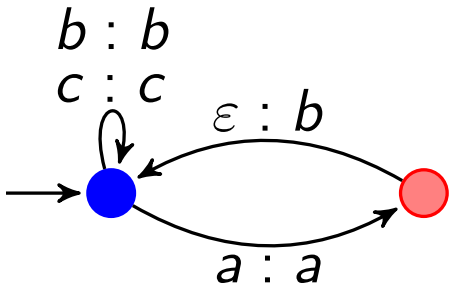
- Добавляет a в начале слова:



- Удаляет конечную b в тех словах, где она есть, и отвергает остальные:

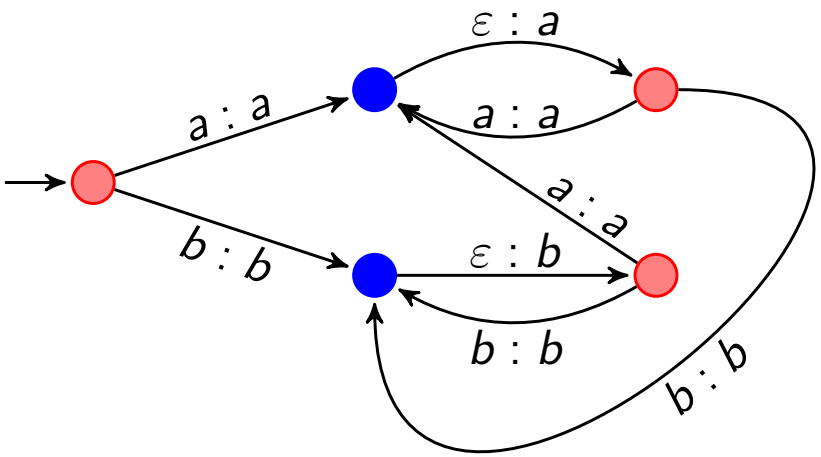


- После каждой a добавляет b:

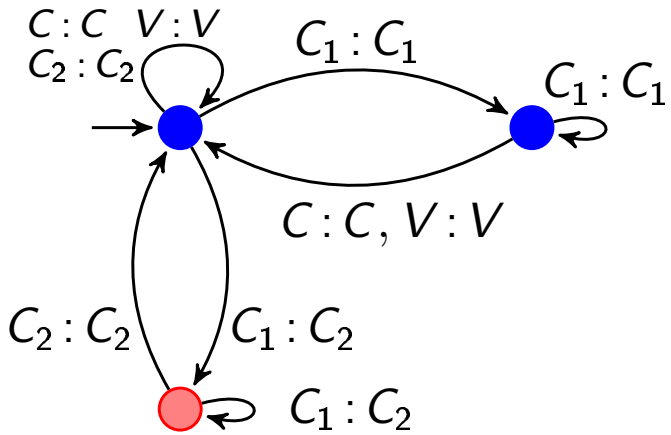


Конечные преобразователи: примеры

- Удваивает все буквы, кроме последней:



- Ретро-ассимилирует C_1 в C_2 (в последовательности C_1, идущей перед C_2, все буквы заменяются на C_2)



Свойства конечных автоматов

- Конечные преобразователи задают в точности регулярные преобразования (аналог теоремы Клини).
- Замкнутость относительно операций:

Операция	Автоматы	Преобразователи
Конкатенация	Да	Да
Итерация	Да	Да
Объединение	Да	Да
Пересечение	Да	Нет
Дополнение	Да	Нет

Свойства конечных преобразователей

- Конечные преобразования также замкнуты относительно
  - Композиции ($\circ$).
  - Приоритетного объединения ($\cup_p$).

$$(T_1 \cup_p T_2)(u) = \begin{cases} T_1(u), & T_1(u) \neq \varnothing, \\ T_2(u), & T_1(u) = \varnothing. \end{cases}$$

  - Обращения $((\cdot)^{-1})$: $\phi^{-1} = \{\langle y, x\rangle \mid \langle x, y\rangle \in \phi\}$.
- Применения операций:
  - Композиция: последовательное применение преобразований,
  - Обращение: переход от синтеза к анализу и наоборот,
  - Приоритетное объединение: отдельная обработка нерегулярных форм.

Множественное число существительного в английском

Пример.

Опишите преобразователь, преобразующий форму единственного числа существительного в форму множественного для английского языка.

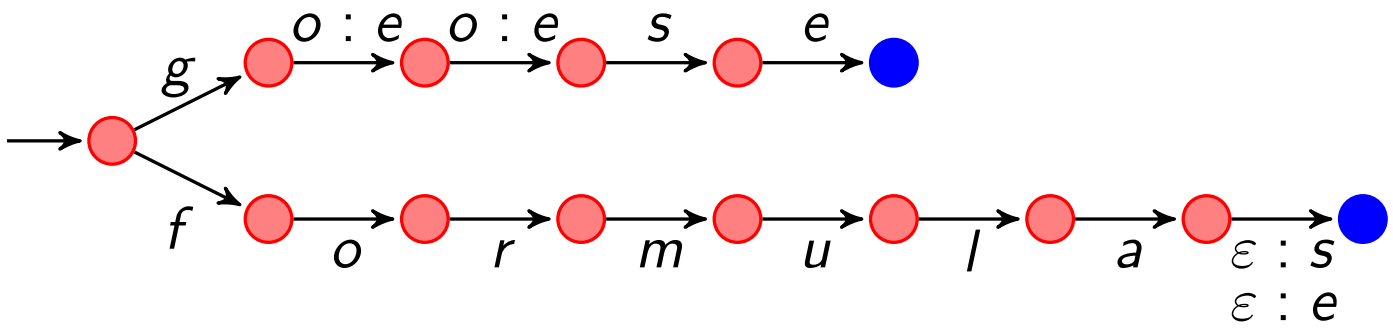
- *torch* $\leftrightarrows$ *torches*
- *monarch* $\leftrightarrows$ *monarchs*
- *ally* $\leftrightarrows$ *allies*
- *play* $\leftrightarrows$ *plays*
- *goose* $\leftrightarrows$ *geese*
- *formula* $\leftrightarrows$ *formulas/formulae*

Пример: множественное число в английском

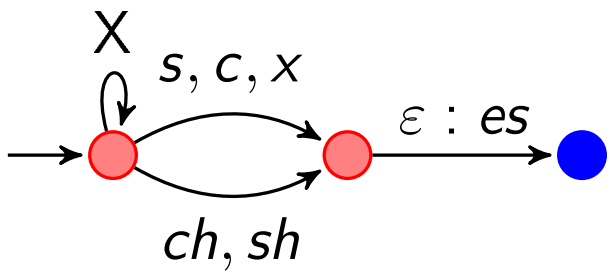
Пример.

Описать преобразователь, строящий форму множественного числа существительного в английском фзыке.

- Отдельный преобразователь T_{exc} для исключений:

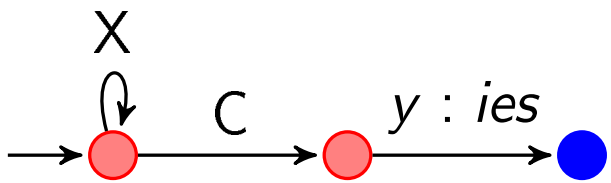


- T_{sib}: добавляет -*es* после финальной шипящей (X — произвольный символ):



Пример: множественное число в английском

- Преобазователь T_{exc} для исключений.
- Преобразователь T_{sib} для добавления *es*.
- Преобразователь T_{Cy} для замены *y* на *-ies* после согласной.



- T_s — добавляет s в конце.
- $T_{exc,sib}$ — добавляет s к исключениям на *-arch* и отвергает остальные слова (для *monarchs*, *tetrarchs*, ...).
- Общее решение:

$$T_{exc} \cup_p T_{exc,sib} \cup_p T_{sib} \cup_p T_{Cy} \cup_p T_s$$

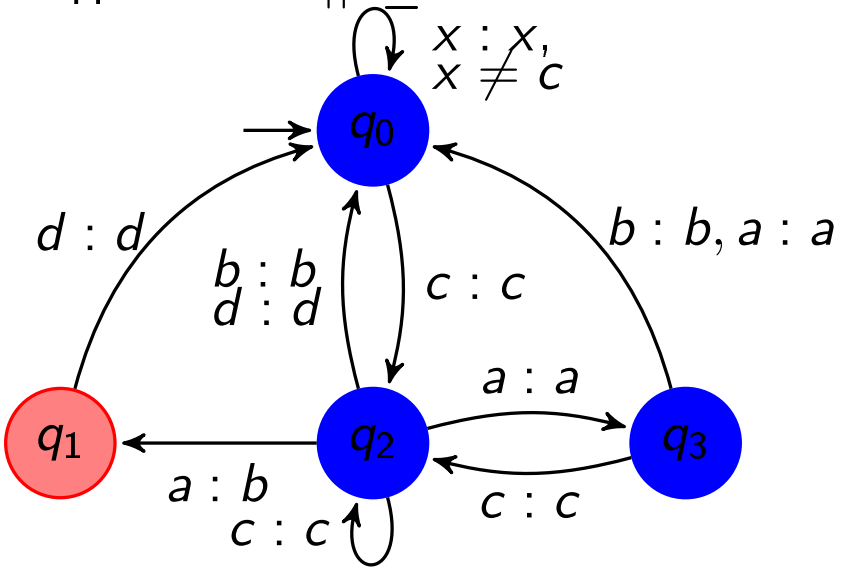
Контекстная замена

- Самый частый тип преобразования — контекстная замена:

$$X \to Y \,||\, U\_V$$

"Заменить X на Y, если слева стоит U, а справа V."

- В простейшем случае X, Y, U, V — буквы.
- Преобразователь для $a \to b||c\_d$:



- В общем случае X, Y, U, V — произвольные регулярные выражения.

Пример: снова множественное число

- Наша модель для английского лингвистически неадекватна.
- Нет отдельных окончаний *-es*, *-ies*, *-s*, они алломорфы окончания *-s*.
- Нужно сначала присоединить прототипическое окончание, потом моделировать контекстные явления.
- Преобразователи для правил:
  - T_s: добавить $!s$ в конце слова ($!$ — маркер границы) $\varepsilon \rightarrow !s \,||\, \_\$$ ($\$$ — маркер конца слова).
  - T_{sib}: добавить e после $!$ и после шипящей $\varepsilon \rightarrow e \,||\, (s|z|x|sh|ch)\_!$.
  - T_y: заменить y на ie перед маркером границы и после согласной $y \rightarrow ie \,||\, C\_!$.
  - $T_{exc,sib}$: ничего не делать со словами с *arch* на конце (возможно, не все такие слова подойдут). $?^* V ?^* arch!s$
  - T_c: удалить маркер морфемной границы $! \rightarrow \varepsilon$.
- Финальный преобразователь строится с помощью композиции:

$$T_{exc} \cup_p (T_s \circ (T_{exc,sib} \cup_p T_{sib}) \circ T_y \circ T_c)$$

Примеры вычисления

- Исключения: $mouse \mapsto mice$, $cactus \mapsto cactuses, cacti$
  — обрабатывается в T_{exc} и не идёт в основную ветку.
- Регулярные формы обрабатываются в ветке
  $T_s \circ (T_{exc,sib} \cup_p T_{sib}) \circ T_y \circ T_c$:
  - $day \to day!s \to day!s \to day!s \to days$
  - $rally \to rally!s \to rally!s \to rallie!s \to rallies$
  - $witch \to witch!s \to witche!s \to witche!s \to witches$
- Частичное исключение *monarch* обрабатывается веткой
  $T_s \circ T_{exc,sib} \circ T_y \circ T_c$:
  $monarch \to monarch!s \to monarch!s \to monarch!s \to monarchs$.

Инфинитив пассивного залога в турецком

Инфинитив пассивного залога

Построить конечный преобразователь, преобразующий турецкий глагольный инфинитив в инфинитив пассивного залога.

- Пассив образуется вставкой суффикса перед *-mek/-mak*.
- Суффикс пассива: *-n* после гласной, *-In* после *l* и *-Il* иначе.
- I: *ı* после *a*, *ı*; *u* после *u*, *o*; *i* после *e*, *i*; *ü* после *ü*, *ö*.

Инфинитив	Инфинитив пассива
aktarmak “*перемещать*”	*aktarılmak*
silmek “*удалять*”	*silinmek*
büyümek “*увеличивать*”	*büyünmek*
durmak “*останавливать*”	*durulmak*
bilmek “*знать*”	*bilinmek*

Инфинитив пассивного залога в турецком

Инфинитив пассивного залога

Построить конечный преобразователь, преобразующий турецкий глагольный инфинитив в инфинитив пассивного залога.

- Пассив образуется вставкой суффикса перед *-mek/-mak*.
- Суффикс пассива: *-n* после гласной, *-In* после *l* и *-Il* иначе.
- I: *ı* после *a*, *ı*; *u* после *u*, *o*; *i* после *e*, *i*; *ü* после *ü*, *ö*.

- T_{mark}: вставить ! перед *-mak/-mek*: $\varepsilon \rightarrow ! \,||\, \_m(a|e)k\$$.
- Заменить маркер подходящим суффиксом:
  - *-n* после гласной (T_V): $! \rightarrow n \,||\, V \_ \$$,
  - *-In* после *l* (T_l): $! \rightarrow \mathrm{I}n \,||\, l \_ \$$,
  - *-Il* по умолчанию (T_{def}): $! \rightarrow \mathrm{I}l \,||\, \_$,
- Соединить всё вместе $T_{suf} = T_V \circ T_l \circ T_{def}$.

Инфинитив пассивного залога в турецком

Инфинитив пассивного залога

Построить конечный преобразователь, преобразующий турецкий глагольный инфинитив в инфинитив пассивного залога.

- Пассив образуется вставкой суффикса перед *-mek/-mak*.
- Суффикс пассива: *-n* после гласной, *-In* после *l* и *-Il* иначе.
- I: *ı* после *a*, *ı*; *u* после *u*, *o*; *i* после *e*, *i*; *ü* после *ü*, *ö*.

- T_{mark} — вставляет ! перед *-mak/-mek*.
- T_{suf} заменяет маркер на подходящий суффикс.
- T_{fill} заполняет гласную суффикса: $T_{fill} = T_\imath \circ T_u \circ T_i \circ T_U$, where
  - $T_\imath$ проверяет условие для $\imath$: $\mathrm{I} \to \imath \,||\, (a|\imath)C^*\_$.
  - T_u для u: $\mathrm{I} \to u \,||\, (u|o)C^*\_$.
  - T_i для i: $\mathrm{I} \to i \,||\, (e|i)C^*\_$.
  - T_U для $\ddot{u}$: $\mathrm{I} \to \ddot{u} \,||\, (\ddot{u}|\ddot{o})C^*\_$.
- Финальный ответ:

$$T_{mark} \circ T_{suf} \circ T_{fill}$$

Глагольные формы в языке йоулумни (Калифорния)

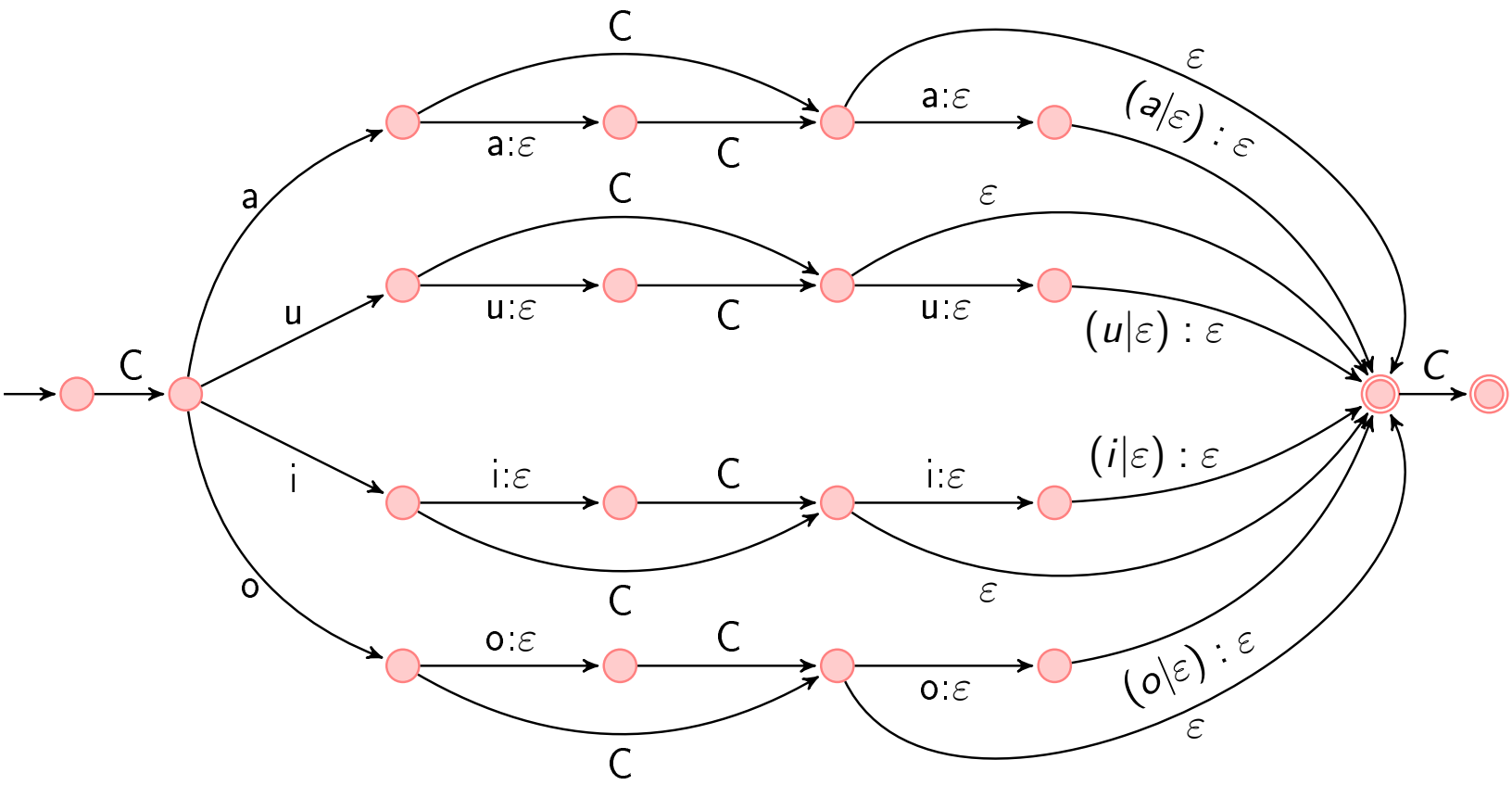
основа	герундий	дуратив
caw “кричать”	caw-inay	cawaa-ʔaa-n
cuum “разрушать”	cum-inay	cumuu-ʔaa-n
hoyoo “называть”	hoy-inay	hoyoo-ʔaa-n
diiyl “охранять”	diyl-inay	diyiil-ʔaa-n
ʔilk “петь”	ʔilk-inay	ʔiliik-ʔaa-n
hiwiit “гулять”	hiwt-inay	hiwiit-ʔaa-n

Глагольные формы в языке йоулумни

Если основа имела вид $\alpha_1 V(V)\alpha_2(V)(V)\alpha_3$, где $\alpha_1, \alpha_2 \in C$, $\alpha_3 \in \{C, \varepsilon\}$, то основа герундия имеет вид $\alpha_1 V \alpha_2 \alpha_3$, а основа дуратива — $\alpha_1 V \alpha_2 VV \alpha_3$.

Преобразователи для глагольных форм в языке йоулумни

- Основа герундия:



Преобразователи для глагольных форм в языке йоулумни

- Основа дуратива: