Морфологический анализатор DIRETRA: больше, чем глосса

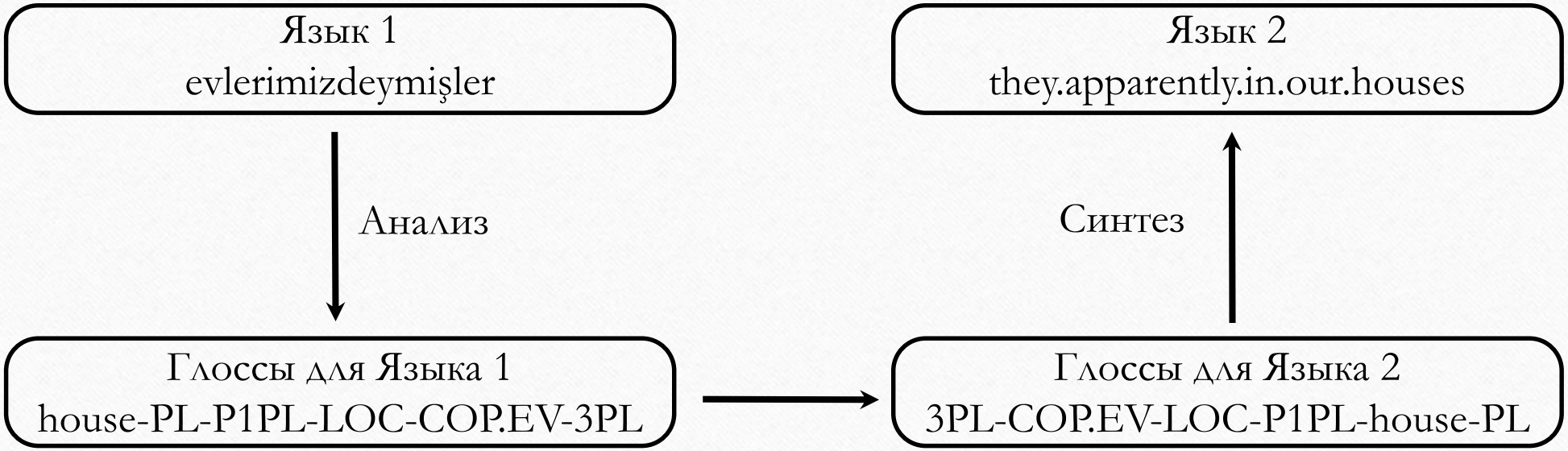
Алёна Аксёнова, Марина Ермолаева

25 октября 2014

The Zen of Diretra

- Diretra – подстрочник с морфологией.
- Diretra ≠ переводчик.
- Лингвистические обобщения – наши друзья.
- Если это возможно, это нужно учесть.
- Сперва общее, затем частное.
- Ошибки недопустимы.
- Простота – одно из лучших качеств человека © Конфуций
  ... и программы.

Схема работы



Анализ

- Морфологический парсер выполняет анализ (глоссирование) исходной словоформы Языка 1;
- На выходе – все возможные варианты разбора словоформы.

- вход:

evlerimizdeymişler

- выход:

ev-ler-imiz-de-ymiş-ler
house-PL-P1PL-LOC-COP.EV-3PL

Разметка соответствий

- Каждому допустимому морфологическому разбору словоформы для Языка 1 ставится в соответствие глосса для Языка 2
- Глоссы могут означать как морфологическую, так и синтаксическую единицу.

- вход:

house-PL-P1PL-LOC-COP.EV-3PL

- выход:

3PL-COP.EV-LOC-P1PL-house-PL

Синтез

- Синтезируется словоформа на Языке 2, соответствующая исходной.

- вход:

3PL-COP.EV-LOC-P1PL-house-PL

- выход:

they.apparently.in.our.houses

Turkish Challenges: фонология

- Гармония гласных:
  - палатальная
  - губная
- Отсутствие гармонии:
  - внутри основ и сложных слов
  - в ряде заимствованных слов с заднерядным гласным последнего слога.

baş-**ın**	kol-**un**
голова-P2SG	рука-P2SG
ev-**in**	göz-**ün**
дом-P2SG	глаз-P2SG
hal-**in**	
состояние-P2SG	

Turkish Challenges: фонология

- Варианты аффиксов:
  - звонкий/глухой
  - гласный/согласный

oda-**da**	sokak-**ta**
комната-LOC	улица-LOC
ev-**i**	elbise-**si**
дом-P3SG	платье-P3SG

Turkish Challenges: изменения в корнях

- Несколько классов основ с особыми свойствами:
  - озвончение
  - чередование гласного с нулем
  - удвоение согласного

dolap шкаф	dolab-a шкаф-DAT
şehir город	şehr-i город-ACC
hak-lar право-PL	hakk-ın право-P2SG

Turkish Challenges: именная морфология

- Основные морфологические категории имени:
  - число;
  - посессивность;
  - падеж.

STEM	NUM	POSS	CASE
çocuk	-lar	-ın	-a
ребенок	-PL	-P2SG	-DAT

твоим детям

Turkish Challenges: именная морфология

- Суффикс -ki присоединяется к именам в генитиве или локативе;
- Формы на -ki могут затем принимать именные суффиксы;
- LOC-ki и GEN-ki имеют разные свойства (Hankamer 2004);
- Цикличность: -ki может присоединяться к локативным и генитивным формам, уже содержащим -ki.

raf-ta-ki
полка-LOC-KI1
тот, что находится на полке

Hasan-ın-ki
Хасан-GEN-KI2
тот, что принадлежит Хасану

ev-de-ki-ler-in-ki
дом-LOC-KI1-PL-GEN-KI2
тот, что принадлежит тем, что находятся в доме

Turkish Challenges: глагольная морфология

- Именные словоформы могут образовывать:
  - предикаты, присоединяя связки и глагольные лично-числовые показатели;
  - адвербиальные формы, присоединяя соответствующие суффиксы.

STEM	NUM	POSS	CASE	COPULA	PERS+NUM
ev	-ler	-imiz	-de	-ymiş	-ler
дом	-PL	-P1PL	-LOC	-COP.EV	-3PL

По-видимому, они (были) в наших домах.

STEM	CASE	ADV
sokak	-ta	-yken
улица	-LOC	-while

будучи на улице

Turkish Challenges: сложные слова

- Композиты – сложные слова, образованные соположением основ;
- В турецком языке композиты вида «N+N» или «Adj+N» очень продуктивны;
- Список композитов, пишущихся слитно, открыт.

kız-arkadaş
девушка-друг
(любимая) девушка

sarı-humma
желтый-лихорадка
желтая лихорадка

Парсер: задачи

- Возможность адаптации для работы с разными языками;
- Хорошие результаты при ограниченных ресурсах:
  - анализ морфологии при неизвестной основе
- Обработка сложных случаев:
  - основы с особыми свойствами
  - рекурсивные показатели (-ki)
  - композиты

Парсер: подходы

- Словарь корней и словарь суффиксов:
  - неагглютинативные языки
  - конечный список возможных суффиксов
  - Segalovich 2003
- Конечные автоматы:
  - подходят для агглютинативных языков
  - набор возможных суффиксов бесконечен
  - Hankamer 1986, Eryiğit & Adalı 2004, Sak et al. 2009, Çöltekin 2010, Şahin et al. 2013…

Парсер: описание

- За основную единицу принимается слот – часть цепочки аффиксов, внутри которой порядок следования морфем жестко фиксирован;
- Для каждого слота описываются последовательности категорий, которые могут его заполнять:

Категория	**Последовательности**	**Слот**
Число	-NUM-POSS-CASE	Именное словоизменение
Посессивность		
Падеж		
Вопросительная клитика	-(Q)-COP.PRS-PERS -(Q)-COP.PST-PERS -ADV …	Глагольное словоизменение
Связка		
Лично-числовые суффиксы		
Адвербиальные суффиксы		

Парсер: описание

- Набор слотов фиксирован, но…
- Количество и порядок следования категорий внутри слота можно легко изменить.

№	Категория	№	Слот
1	Основа	1	Основа-модификатор
		2	Основа
2	Число	3	Именное словоизменение
3	Посессивность		
4	Падеж		
5	Показатель -ki	4	«Петля» в именном словоименении
6	Вопросительная клитика	5	Глагольное словоизменение
7	Связка		
8	Лично-числовые суффиксы		
9	Адвербиальные суффиксы		

Парсер: данные о языке

- Набор последовательностей категорий для каждого слота;
- Данные о фонологии: инвентарь фонем, описание гармонии гласных;
- Список аффиксов:
  - в список вносятся все алломорфы с ярлыками
  - для каждого алломорфа указывается набор (мор)фонологических ограничений, если они есть
- Словарь основ (опционально)

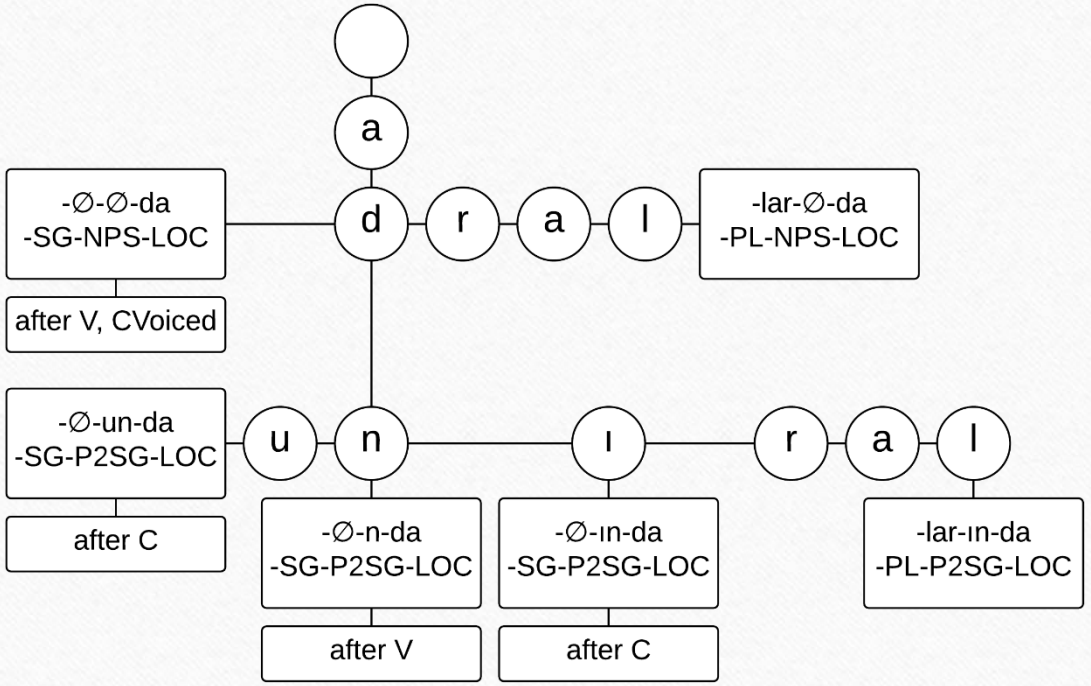
Парсер: представление данных

- Для каждого слота строятся все возможные цепочки аффиксов;
- Используется информация о порядке следования категорий в слоте, гармонии и ограничениях на контекст.

Цепочка аффиксов	**Глосса**	**Ограничения**
-∅-∅-da	-SG-NPS-LOC	после гласных, звонких согласных
-∅-∅-de	-SG-NPS-LOC	после гласных, звонких согласных
-∅-n-da	-SG-P2SG-LOC	после гласных
-∅-n-de	-SG-P2SG-LOC	после гласных
-lar-∅-da	-PL-NPS-LOC	
-ler-∅-de	-PL-NPS-LOC	
-∅-ın-da	-SG-P2SG-LOC	после согласных
-∅-un-da	-SG-P2SG-LOC	после согласных

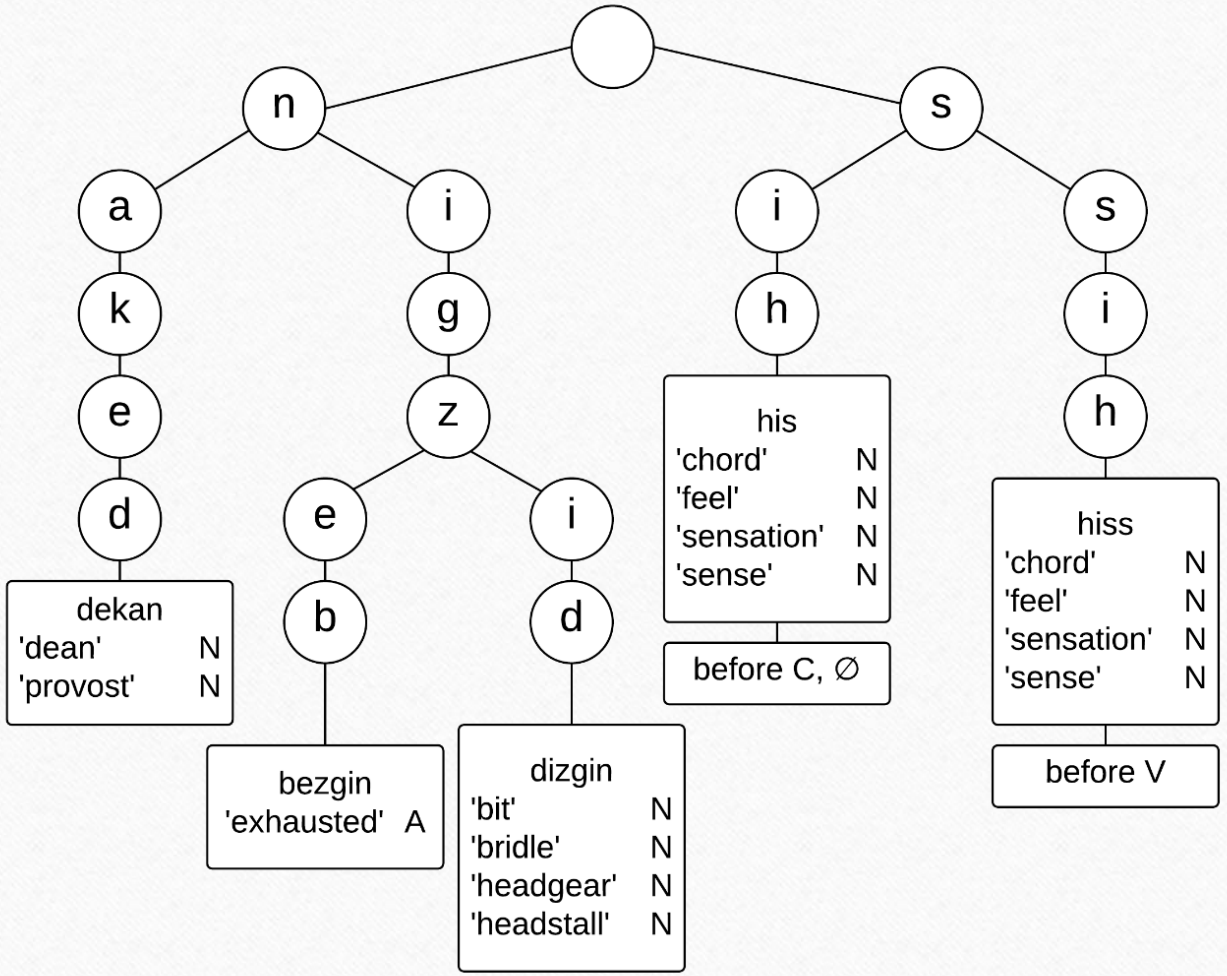
Парсер: представление данных

- Списки последовательностей преобразуются в буквенные деревья;
- Цепочки аффиксов хранятся в обратном порядке.



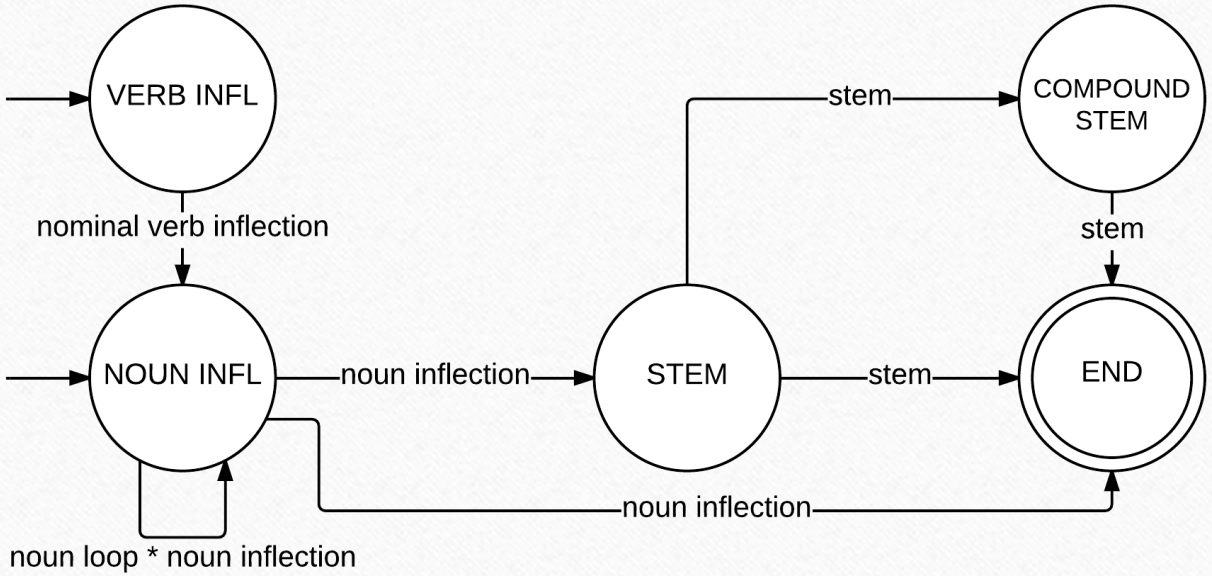
Парсер: представление данных

- Аналогичные деревья строятся для лексикона;
- Основы с несколькими фонологическими вариантами рассматриваются как отдельные вхождения.



Парсер: алгоритм анализа

- Переходы между слотами осуществляются с помощью конечного автомата, в котором каждый переход соответствует одному слоту.



Парсер: алгоритм анализа

- Двигаясь справа налево, найти все возможные цепочки аффиксов, которые могут соответствовать словоформе на входе;
- Для каждого гипотетического разбора попытаться найти основу в словаре, которая бы соответствовала неотглоссированной части;
- Если основа нашлась и «лишних» символов не осталось: выдать все такие разборы;
- Если основа нашлась, но слева остались неразобранные символы: попытаться найти в словаре соответствующую им основу; выдать все разборы со сложной основой;
- Если ни для одного разбора не удается найти простую или сложную основу: выдать все гипотетические разборы.

Парсер: примеры

- Вход: *adamdɪ*
- Гипотетическая основа:
  - *adam «человек»*: найдено
  - *ada «остров»*: найдено
  - *adamd*: не найдено
- Решение: простая основа
- Результат:

adam-∅-∅-∅-dɪ-∅
man-SG-NPS-NOM-COP.PST-3

ada-∅-m-∅-dɪ-∅
island-SG-P1SG-NOM-COP.PST-3

Парсер: примеры

- Вход: *kızarkadaş*
- Гипотетическая основа:
  - *kızarkadaş*
    - *adaş «тезка»*: найдено
      - *kızark*: не найдено
    - *arkadaş «друг»*: найдено
      - *kız «девушка»*: найдено
- Решение: сложная основа

- Результат:

kız-arkadas-∅-∅-∅
girl-friend-SG-NPS-NOM

kız-arkadas-∅-∅-∅-∅-∅
girl-friend-SG-NPS-NOM-COP.PRS-3

Парсер: примеры

- Вход: *fefe*
- Гипотетическая основа:
  - *fefe*: не найдено
  - *fef*: не найдено
- Решение: неизвестная основа

- Результат:

fef-∅-∅-e
FEF-SG-NPS-DAT

fef-∅-∅-e-∅-∅
FEF-SG-NPS-DAT-COP.PRS-3

fefe-∅-∅-∅
FEFE-SG-NPS-NOM

fefe-∅-∅-∅-∅-∅
FEFE-SG-NPS-NOM-COP.PRS-3

Разметка соответствий: основная идея

- То, что в одних языках – синтаксис, в других – морфология;
- Тюркские языки: богатая морфология;
- Английский язык: бедная морфология.

- The Mirror Principle (Baker 1985):

  Morphological derivations must directly reflect syntactic derivations (and vice versa).

Разметка соответствий: основная идея

- Для языка с бедной морфологией мы “отзеркаливаем” последовательность аффиксов языка с богатой морфологией;
- На месте остаются те аффиксы, которые могут быть морфемами в Языке 2 и относятся к основе.

ada-∅-m-∅-∅-lar
island-SG-P1SG-NOM-COP.PRS-3PL
3PL-COP.PRS-NOM-P1SG-island-SG
they.are.my.island

çocuk-lar-∅-ın-ki-ler-∅-∅
child-PL-NPS-GEN-KI2-PL-NPS-NOM
NOM-NPS-PL-KI2-GEN-NPS-child-PL
ones.owned.by.children

Разметка соответствий: вторичное

- Если в Языке 2 есть передвижение вспомогательного глагола при общем вопросе:
- При наличии вопросительного маркера происходит передвижение временных связок в крайнюю левую позицию.

çocuk-lar-Ø-Ø-mı-Ø-yız
child-PL-NPS-NOM-Q-COP.PRS-1PL
COP.PRS-1PL-NOM-NPS-child-PL-Q
are.we.children.?

Синтез: правила порождения

- 5 типов правил:
  - замена (“DAT” → “to”, “ACC” → “the”)
  - контекстная замена (“COP.PRS” → “am” if 1SG, → “is” if 3SG, → “are” elsewhere)
  - порождение с учётом фонологии (“PL” → “ies” if __Cy, → “PL” → “ves” if __f etc.)
  - порождение по списку (“deer” + “PL” → “deer”)
  - статистический подсчёт и замена (“LOC” → “in” or “on” or “at”)

Синтез: правила замены

- “DAT” → “to”
- “GEN” → “of.the” при основе, “by” при KI
- “ACC” → “the” (DOM в тюркских языках, Lyutikova&Pereltsvaig (2013))
- “1SG” → “I”, “1PL” → “we”, ...

arkadaş-Ø-Ø-ı
friend-SG-NPS-ACC
ACC-NPS-friend-SG
the.friend

arkadaş-Ø-ım-ı
friend-SG-P1SG-ACC
ACC-P1SG-friend-SG
my.friend

Синтез: контекстная замена

- Так происходит согласование;
- В случае “3”, т.е. показателя, специфицированного только по лицу, мы выбираем английский аналог с минимальной спецификацией.

çocuk-Ø-Ø-Ø-mı-Ø-yım
child-SG-NPS-NOM-Q-COP.PRS-1SG
COP.PRS-1SG-NOM-NPS-child-SG-Q
am.I.child.?

adam-Ø-Ø-Ø-Ø-Ø
man-SG-NPS-NOM-COP.PRS-3
3-COP.PRS-NOM-NPS-man-SG
it.is.man

Синтез: порождение по списку

- Список неправильных форм множественного числа.

geyik-ler-Ø-Ø
deer-PL-NPS-NOM
NOM-NPS-deer-PL
deer

eksen-ler-Ø-Ø
axis-PL-NPS-NOM
NOM-NPS-axis-PL
axes

Синтез: порождение с учетом фонологии

- Используется для образования правильных форм множественного числа.

karı-lar-Ø-Ø
wife-PL-NPS-NOM
NOM-NPS-wife-PL
wives

sihirbaz-lar-Ø-Ø
witch-PL-NPS-NOM
NOM-NPS-witch-PL
witches

Синтез: статистический подсчет и замена

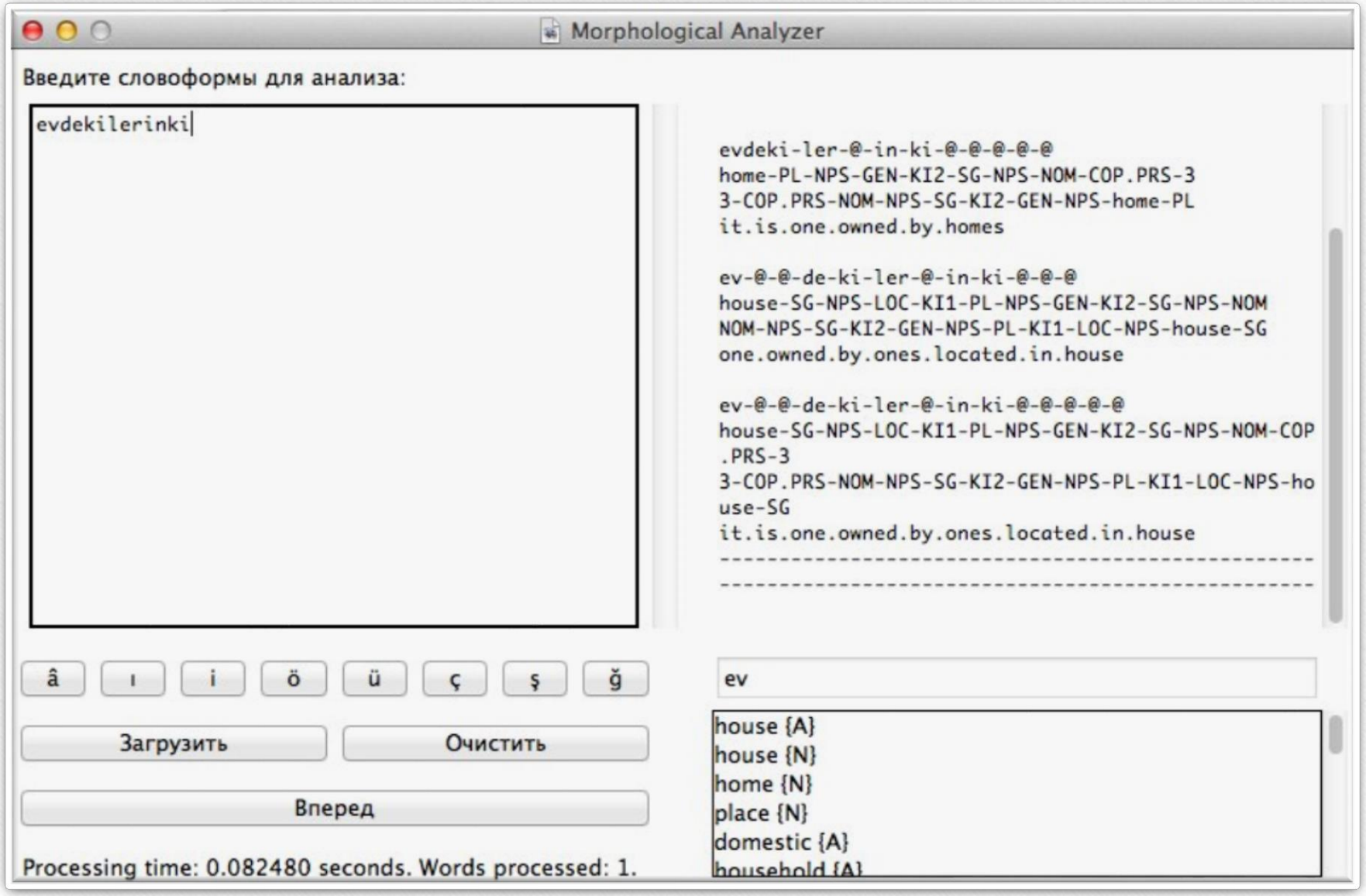
- Для каждого английского существительного имеется локативный предлог (in/at/on), который с ним употребляется чаще всего;
- На него заменяется “LOC”.

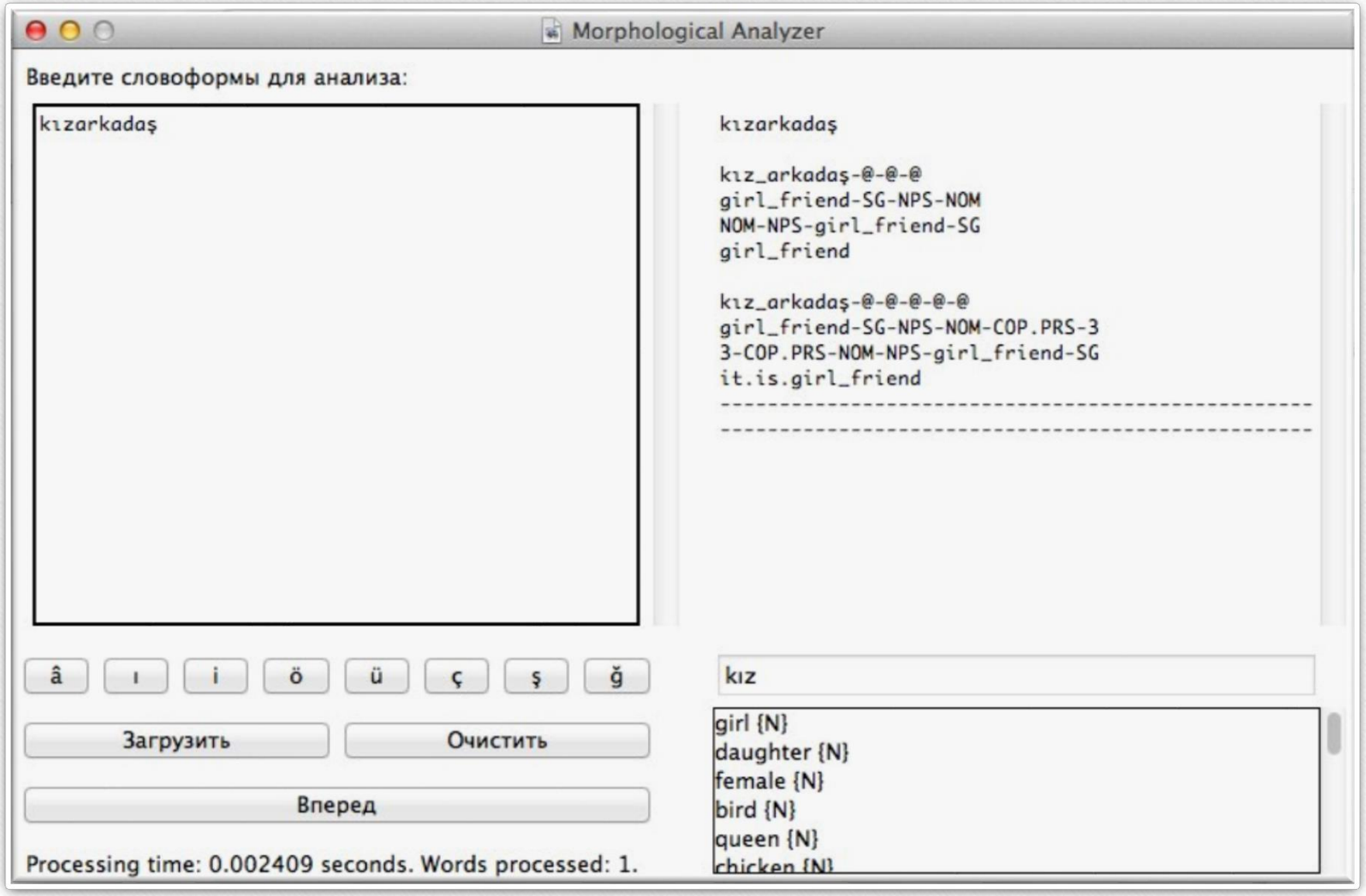
ev-∅-∅-de
house-SG-NPS-LOC
LOC-NPS-house-SG
in.house

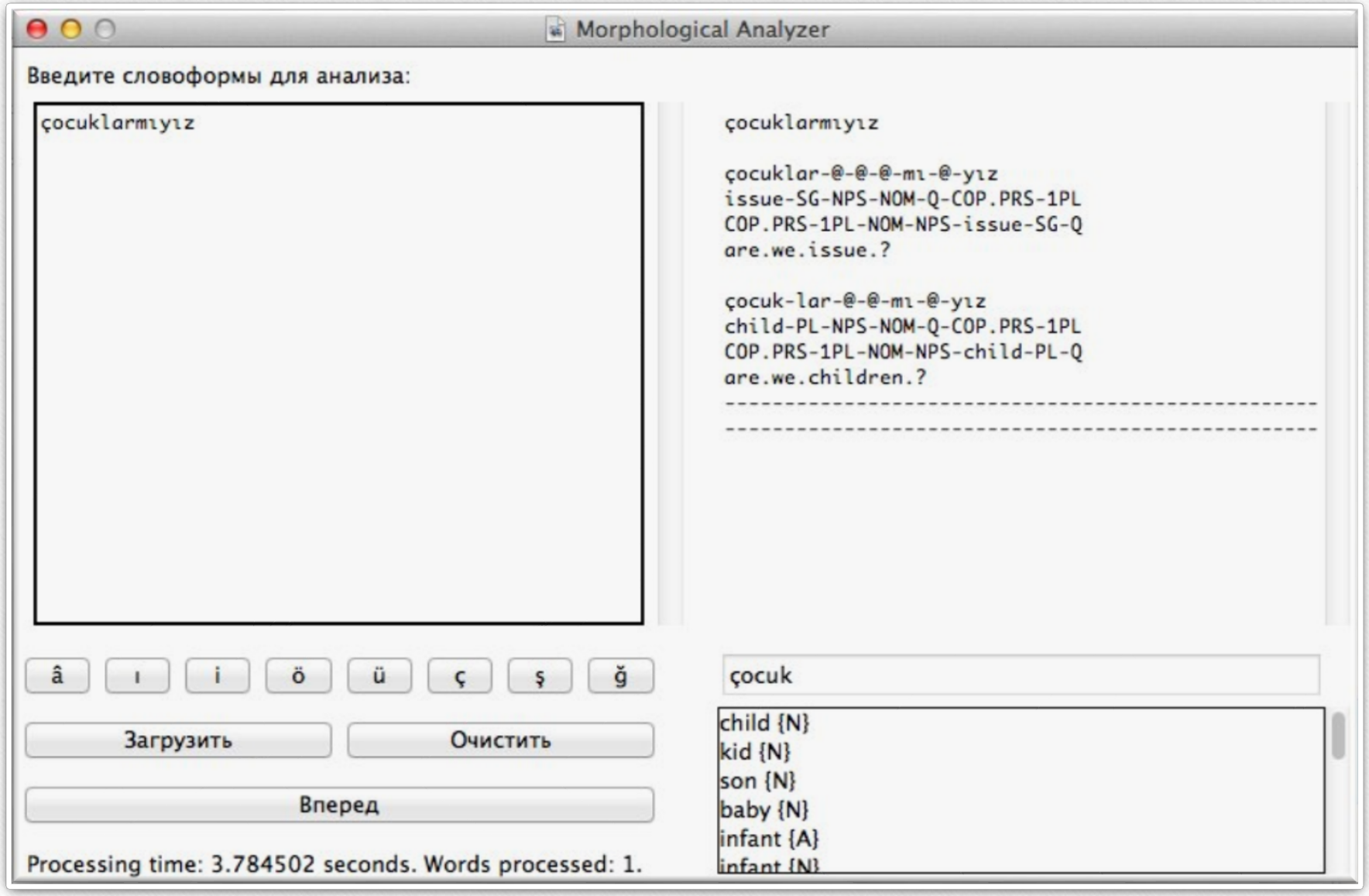
p(in) = 0.77774

okul-∅-∅-da
school-SG-NPS-LOC
LOC-NPS-school-SG
at.school

p(at) = 0.999994


Morphological Analyzer
Введите словоформы для анализа:
evdekilerinki
evdeki-ler-@-in-ki-@-@-@-@-@
home-PL-NPS-GEN-KI2-SG-NPS-NOM-COP.PRS-3
3-COP.PRS-NOM-NPS-SG-KI2-GEN-NPS-home-PL
it.is.one.owned.by.homes
ev-@-@-de-ki-ler-@-in-ki-@-@-@
house-SG-NPS-LOC-KI1-PL-NPS-GEN-KI2-SG-NPS-NOM
NOM-NPS-SG-KI2-GEN-NPS-PL-KI1-LOC-NPS-house-SG
one.owned.by.ones.located.in.house
ev-@-@-de-ki-ler-@-in-ki-@-@-@-@-@
house-SG-NPS-LOC-KI1-PL-NPS-GEN-KI2-SG-NPS-NOM-COP
.PRS-3
3-COP.PRS-NOM-NPS-SG-KI2-GEN-NPS-PL-KI1-LOC-NPS-ho
use-SG
it.is.one.owned.by.ones.located.in.house
â ı i ö ü ç ş ğ
ev
Загрузить
Очистить
house {A}
house {N}
home {N}
place {N}
domestic {A}
household {A}
Вперед
Processing time: 0.082480 seconds. Words processed: 1.


Morphological Analyzer
Введите словоформы для анализа:
kızarkadaş
kızarkadaş
kız_arkadaş-@-@-@
girl_friend-SG-NPS-NOM
NOM-NPS-girl_friend-SG
girl_friend
kız_arkadaş-@-@-@-@-@
girl_friend-SG-NPS-NOM-COP.PRS-3
3-COP.PRS-NOM-NPS-girl_friend-SG
it.is.girl_friend
â ı i ö ü ç ş ğ
kız
Загрузить
Очистить
Вперед
girl {N}
daughter {N}
female {N}
bird {N}
queen {N}
chicken {N}
Processing time: 0.002409 seconds. Words processed: 1.


Morphological Analyzer
Введите словоформы для анализа:
çocuklarmıyız
çocuklarmıyız
çocuklar-@-@-@-mı-@-yız
issue-SG-NPS-NOM-Q-COP.PRS-1PL
COP.PRS-1PL-NOM-NPS-issue-SG-Q
are.we.issue.?
çocuk-lar-@-@-mı-@-yız
child-PL-NPS-NOM-Q-COP.PRS-1PL
COP.PRS-1PL-NOM-NPS-child-PL-Q
are.we.children.?
â ı i ö ü ç ş ğ
çocuk
Загрузить
Очистить
child {N}
kid {N}
son {N}
baby {N}
infant {A}
Вперед
Processing time: 3.784502 seconds. Words processed: 1.

Перспективы развития

- Вглубь:
  - финитные и нефинитные глагольные формы
  - прочие части речи
  - деривационная морфология
- Вширь:
  - другие тюркские языки
  - другие алтайские языки

Литература

- Baker M. (1985). The Mirror Principle and Morphosyntactic Explanation // Linguistic Inquiry 16. 373-416.
- Çöltekin Ç. (2010). A Freely Available Morphological Analyzer for Turkish. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias, eds., 'LREC', European Language Resources Association.
- Eryiğit G., Adalı E. (2004). An Affix Stripping Morphological Analyzer for Turkish // IASTED International Multi-Conference on Artificial Intelligence and Applications. Innsbruck, Austria, pages 299-304.
- Göksel A., Kerslake C. (2005). Turkish: A Comprehensive Grammar.
- Hankamer J. (1986). Finite state morphology and left-to-right phonology // Proceedings of the Fifth West Coast Conference on Formal Linguistics, Stanford, CA, pages 29-34.

Литература

- Hankamer J. (2004). Why there are two ki's in Turkish // Imer and Dogan, eds., Current Research in Turkish Linguistics, Eastern Mediterranean University Press, 13-25.
- Kornfilt J. (1996). On copular clitic forms in Turkish. ZAS Papers in Linguistics 6, 96-114.
- Kornfilt J. (1997). Turkish. London and New York: Routledge.
- Lewis G. (1967). Turkish Grammar. Oxford: Oxford University Press.
- Lyutikova E., Pereltsvaig A. (2013). Elucidating nominal structure in articleless languages: A case study of Tatar // Proceedings of 39th Berkeley Linguistic Society Meeting. — Berkeley.
- Şahin M., Sulubacak U., Eryiğit G. (2013). Redefinition Of Turkish Morphology Using Flag Diacritics // Proceedings of the Tenth Symposium on Natural Language Processing (SNLP-2013).