# Стратегии усвоения параметра падежа: экспериментально-корпусное исследование

П. В. Гращенков
**(МГУ, ООО "Сбердевайсы»)**

Исследование выполнено в рамках проекта РНФ № 22-18-00037

# Цели, задачи и основные предположения:

1

Цель – попытаться установить оптимальные стратегии, работающие при усвоении падежных граммем и способов их выражения

2

Задача – смоделировать процесс усвоения морфологического падежа при усвоении языка

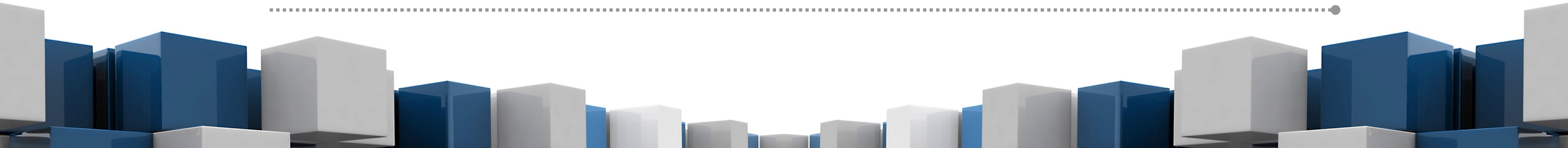
- Предположения:
  - Усваивающий язык ребенок не имеет врожденного знания о к-ве падежей, способе их выражения и т.п.

3

  - В процессе усвоения языка необходимо построить общее представление о падежных граммемах для единиц разных морфологических форм (парадигм)

# Допущения:

1. Изначально доступа к семантике нет или он минимален
2. Входные «обучающие» данные: СинТагРус
3. Формулируется определенный алгоритм обучения, он последовательно применяется ко всем предложениям корпуса
4. Различия в «обучающих данных»



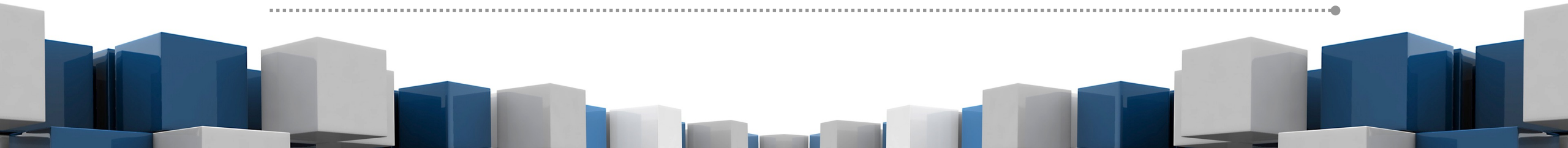
# Что мы знаем об усвоении падежа:

(Babyonyshev 1993), (Цейтлин 2000), (Ladinskaya et al. 2019),…

- Падеж в русском усваивается рано для ед.ч (<20 м.), для мн.ч позже (30 м.)
- Грамматика имен разных склонений (*папа, плакса, доктор*) усваивается с разной скоростью
- Номинатив как падеж усваивается первым в возрасте около 2 л., затем следует аккузатив (~ 3 г.)

# СинТагРус:

Наиболее полный корпус с верифицированной синтаксической разметкой, наилучшие решения обучены на нем.



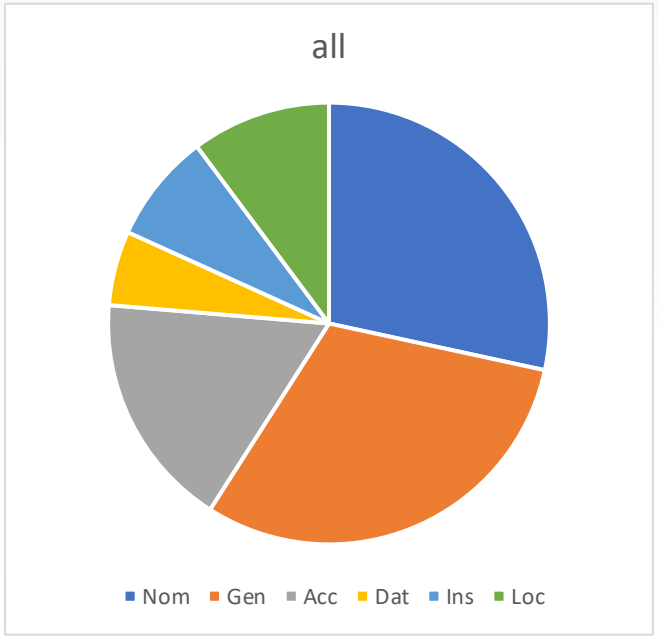
# СинТагРус:

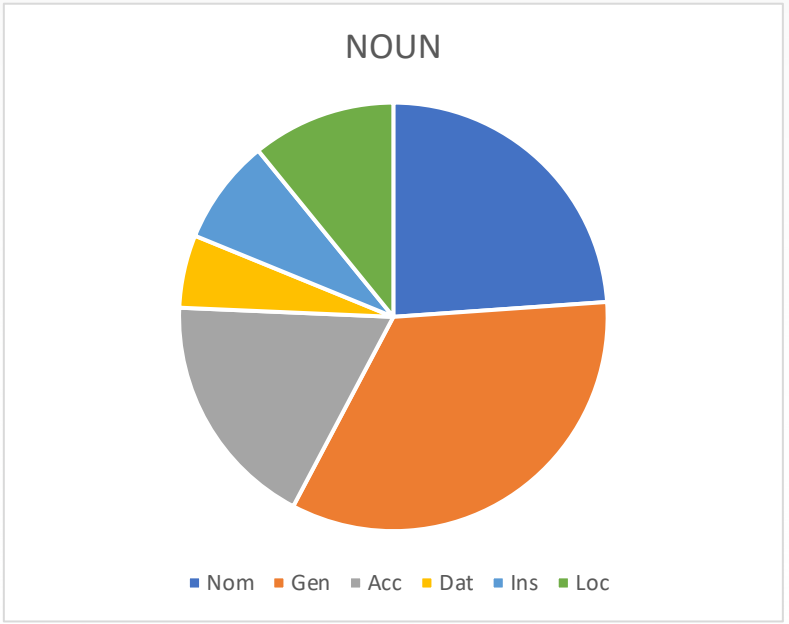
Наиболее полный корпус с верифицированной синтаксической разметкой, наилучшие решения обучены на нем.

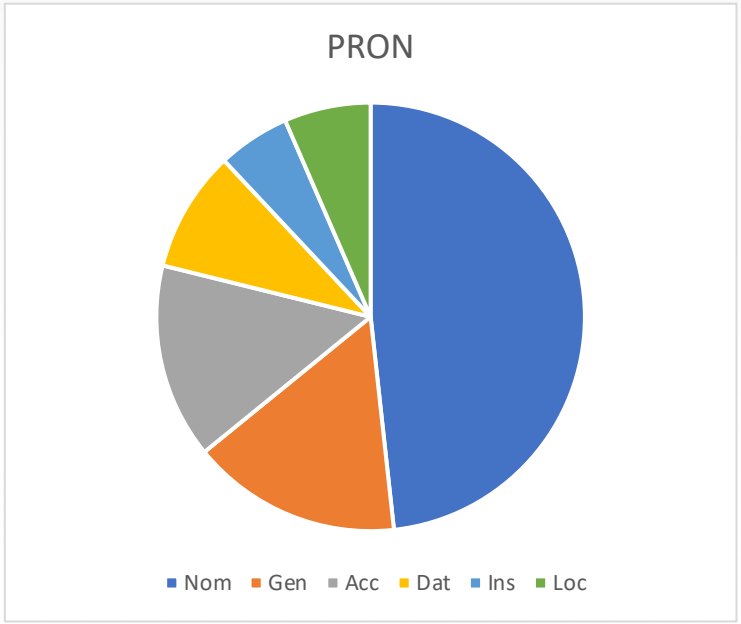
Всего около 100+ т. предложений

| | all | NOUN | PRON |
|---|---|---|---|
| Nom | 142014 | 64678 | 23727 |
| Gen | 153348 | 91627 | 7805 |
| Acc | 86433 | 48580 | 7239 |
| Dat | 27257 | 14880 | 4502 |
| Ins | 40116 | 21549 | 2680 |
| Loc | 51008 | 29400 | 3211 |

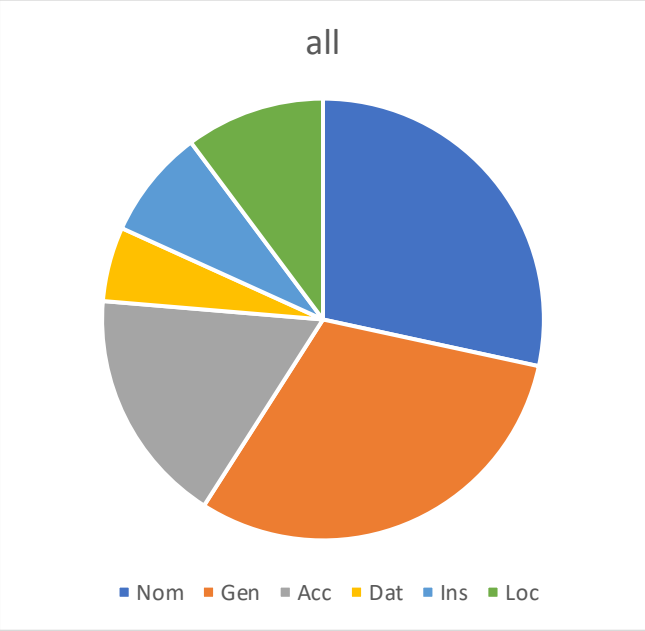
# Падежи в СинТагРусе

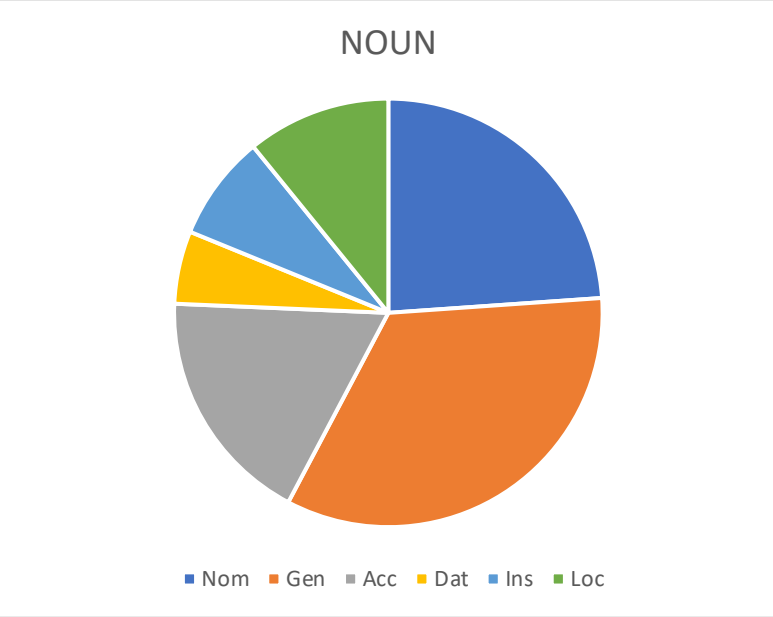


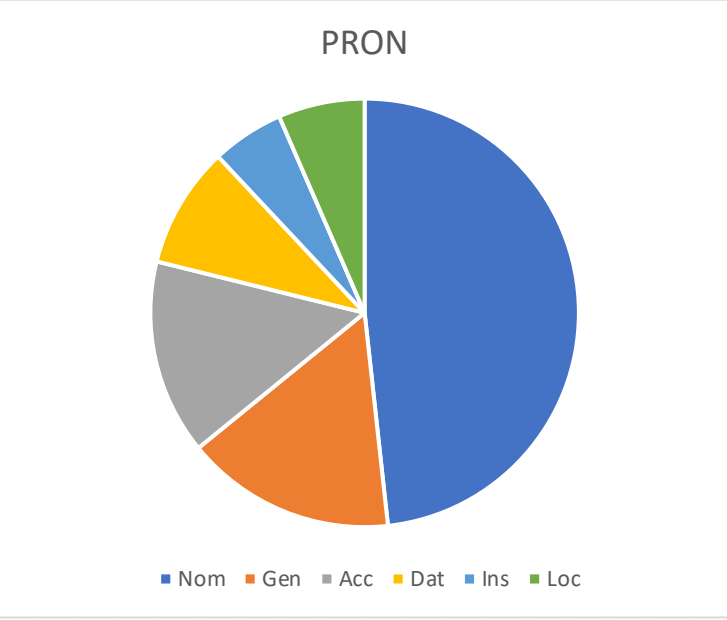



# Падежи в СинТагРусе




Допущение о местоимениях:
Усваиваются ребенком существенно позже имен, не влияют на усвоение набора и средств выражения падежных граммем существительного

# Падежные флексии

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| а/я | мама | коня | дня | | | | |
| С | слон | стол | рыб | | | | |
| о | окно | чудо | | | | | |
| е | море | поле | | вере | | стаде | |
| ы | лбы | полы | папы | | | | |
| и | звери | двери | цели | мели | | твари | |
| у/ю | | тлю | | рабу | | боку | чаю |
| ой | | | | | дамой | | |
| ом/ем | | | | | сном | | |
| ей | | вшей | гусей | | Сашей | | |
| ов | | рабов | снов | | | | |
| ам/ям | | | | пням | | | |
| ами/ями | | | | | котами | | |
| ах/ях | | | | | | словах | |

# Падежные флексии

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| а/я | мама | коня | дня | | | | |
| С | слон | стол | рыб | | | | |
| о | окно | чудо | | | | | |
| е | море | поле | | вере | | стаде | |
| ы | лбы | полы | папы | | | | |
| и | звери | двери | цели | мели | | твари | |
| у/ю | | тлю | | рабу | | боку | чаю |
| ой | | | | | дамой | | |
| ом/ем | | | | | сном | | |
| ей | | вшей | гусей | | Сашей | | |
| ов | | рабов | снов | | | | |
| ам/ям | | | | пням | | | |
| ами/ями | | | | | котами | | |
| ах/ях | | | | | | словах | |

Допущение:
информация об ударении недоступна

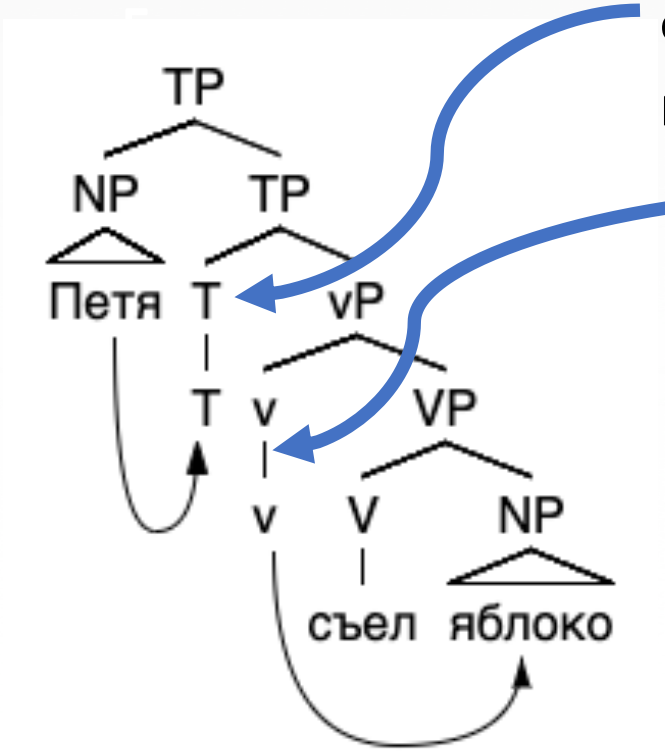
Допущение:
род и число усваиваются отдельно

Допущение:
можно объединять флексии, только если их распределение по падежам полностью совпадает

# Падежные флексии

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| а/я | Nom | Acc | Gen | | | | | 53929 | 0,21 |
| С | Nom | Acc | Gen | | | | | 56999 | 0,22 |
| о | Nom | Acc | | | | | | 5858 | 0,02 |
| е | Nom | Acc | | Dat | | Loc | | 27502 | 0,11 |
| ы | Nom | Acc | Gen | | | | | 17602 | 0,07 |
| и | Nom | Acc | Gen | Dat | | Loc | | 33680 | 0,13 |
| у/ю | | Acc | | Dat | | Loc | GenP | 19971 | 0,08 |
| ой | | | | | Ins | | | 3021 | 0,01 |
| ом/ем | | | | | Ins | | | 10202 | 0,04 |
| ей | | Acc | Gen | | Ins | | | 5112 | 0,02 |
| ов | | Acc | Gen | | | | | 12550 | 0,05 |
| ам/ям | | | | Dat | | | | 3758 | 0,01 |
| ами/ями | | | | | Ins | | | 5210 | 0,02 |
| ах/ях | | | | | | Loc | | 5354 | 0,02 |
| | | | | | | | | | 260748 |

# Падеж в генеративной парадигме, UG




Функциональная вершина T ответственна за присваивание номинатива

Функциональная вершина v ответственна за присваивание аккузатива

Функциональные вершины T и v:

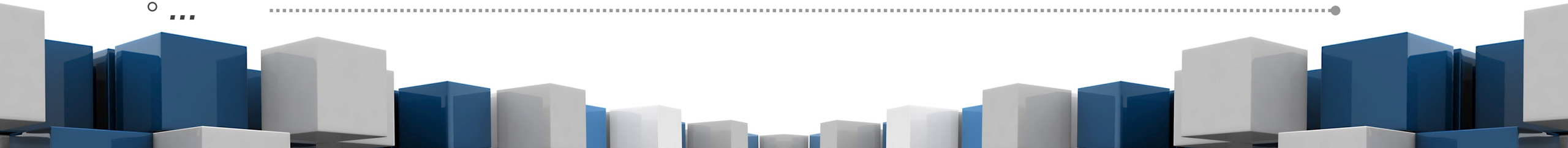
- являются универсальными для всех языков;
- не имеют фонологической реализации и могут установлены лишь по их воздействию на структуру клаузы;
- представляют собой «предустановленное знание»

# Падеж как валентность, Data Driven подход

Падеж – то, что объединяет существительные разных типов в случае их попадания в общий круг дистрибутивных контекстов («падеж по Колмогорову», валентности по Мельчуку и Апресяну, …)

- *есть кашу*
- *любить зверей*
- *целовать маму*
- *мыть чашку*
- *в город*
- *на ужин*
- …

➔ АККУЗАТИВ



# Общая схема изучения разных стратегий

1. Поиск происходит в пределах простого предложения
2. Выбираем в качестве «падежа» некоторую форму имени в соответствии с исследуемой стратегией
3. Записываем количественный результат, «нормированный» по частотности данной флексии в корпусе

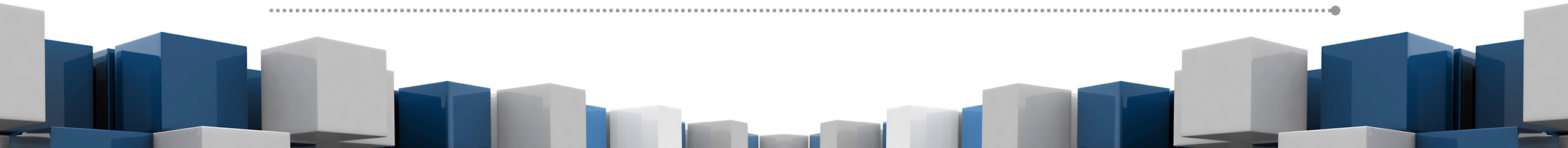
# Что вошло в эксперимент:

5

1. Локатив
2. Номинатив
3. Аккузатив

# Что вошло в эксперимент:

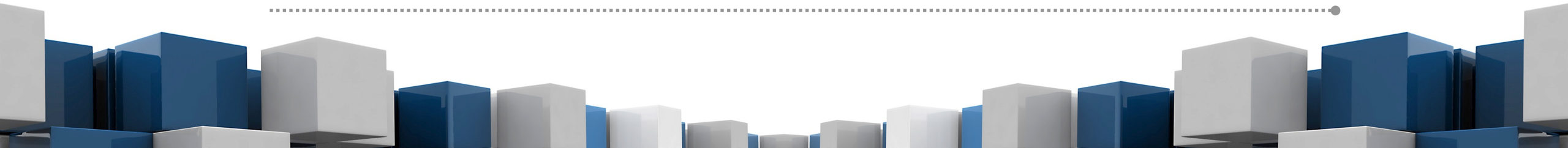
| | Data Driven | UG |
|---|---|---|
| 1. Локатив | 👍 | – |
| 2. Номинатив | 🙃 | 💪 |
| 3. Аккузатив | 🤔 | 😐 |



# Локатив, DD

```
heads = ['о', 'в', 'об', 'обо', 'во', 'на', 'при']
```

1. Запоминаем формы всех существительных слева и справа в окне -5 (-3) head +3 (+5)
2. Смотрим, какие формы наиболее часто попадали в этот контекст

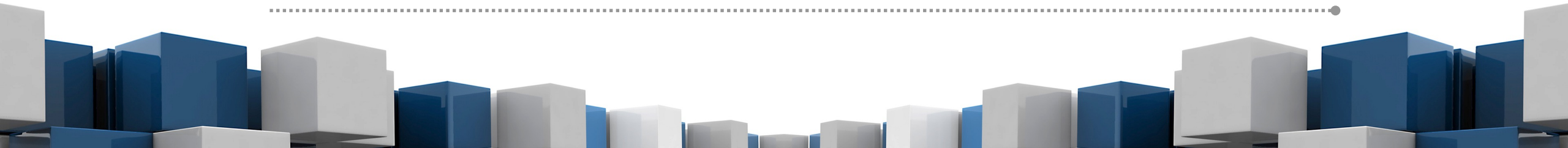


# Локатив, DD

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| а/я | Nom | Acc | Gen | | | | | 107037,757 | | 62038,1907 |
| С | Nom | Acc | Gen | | | | | 102805,133 | | 63440,6439 |
| о | Nom | Acc | | | | | | 108919,487 | | 64986,6985 |
| е | Nom | Acc | | Dat | | Loc | | 232854,733 | | 186520,813 |
| ы | Nom | Acc | Gen | | | | | 112005,206 | | 64527,7973 |
| и | Nom | Acc | Gen | Dat | | Loc | | 139540,438 | | 95256,6328 |
| у/ю | | Acc | | Dat | | Loc | GenP | 132678,443 | | 92974,1379 |
| ой | | | | | Ins | | | 76213,3347 | | 38581,3823 |
| ом/ем | | | | | Ins | | | 86029,9714 | | 48101,1308 |
| ей | | Acc | Gen | | Ins | | | 92067,7113 | | 48762,7324 |
| ов | | Acc | Gen | | | | | 108703,867 | | 61603,0135 |
| ам/ям | | | | Dat | | | | 72992,7877 | | 32957,7701 |
| ами/ями | | | | | Ins | | | 69215,8319 | | 32130,5597 |
| ах/ях | | | | | | Loc | | 340521,109 | | 280910,434 |

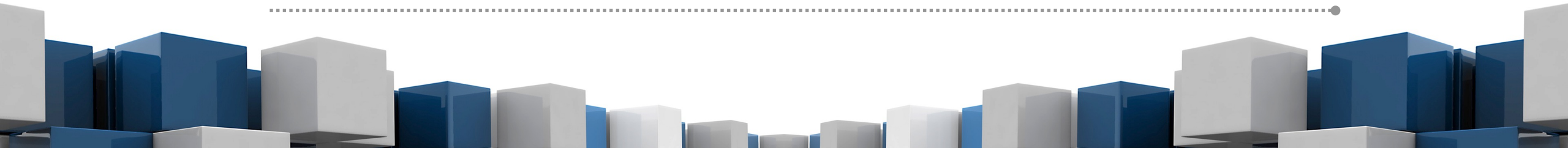
# Номинатив, DD & UG

1. Смотрим на самое левое от финитного глагола (вариант: самое близкое к нему слева) существительное



# Номинатив, UG

2. Смотрим, будут ли «правильные» клетки выигрывать по всем финитным клаузам, демонстрируя функциональную вершину T



# Номинатив, UG

2. Смотрим, будут ли «правильные» клетки выигрывать по всем финитным клаузам, демонстрируя функциональную вершину T: самое близкое слева имя

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| а/я | Nom | Acc | Gen | | | | 33004 |
| с | Nom | Acc | Gen | | | | 41588 |
| о | Nom | Acc | | | | | 53814 |
| е | Nom | Acc | | Dat | | Loc | 51757 |
| ы | Nom | Acc | Gen | | | | 46440 |
| и | Nom | Acc | Gen | Dat | | Loc | 34846 |
| у/ю | | Acc | | Dat | | Loc | 32641 |
| ой | | | | | Ins | | 28483 |
| ом/ем | | | | | Ins | | 35066 |
| ей | | Acc | Gen | | Ins | | 17801 |
| ов | | Acc | Gen | | | | 14004 |
| ам/ям | | | | Dat | | | 43921 |
| ами/ями | | | | | Ins | | 16716 |
| ах/ях | | | | | | Loc | 42760 |

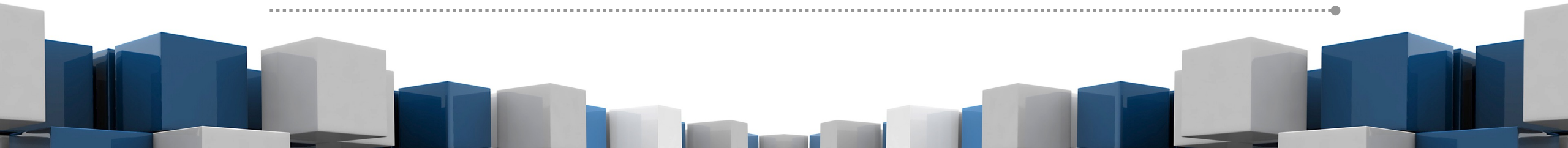
# Номинатив, UG

2. Смотрим, будут ли «правильные» клетки выигрывать по всем финитным клаузам, демонстрируя функциональную вершину T: самое далекое слева имя

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| а/я | Nom | Acc | Gen | | | | | 39917,9567 |
| с | Nom | Acc | Gen | | | | | 39657,2644 |
| о | Nom | Acc | | | | | | 44644,9717 |
| е | Nom | Acc | | Dat | | Loc | | 36833,9023 |
| ы | Nom | Acc | Gen | | | | | 47314,4593 |
| и | Nom | Acc | Gen | Dat | | Loc | | 39158,6515 |
| у/ю | | Acc | | Dat | | Loc | GenP | 25250,9455 |
| ой | | | | | Ins | | | 24685,1797 |
| ом/ем | | | | | Ins | | | 31283,626 |
| ей | | Acc | Gen | | Ins | | | 26778,6972 |
| ов | | Acc | Gen | | | | | 32183,1595 |
| ам/ям | | | | Dat | | | | 32888,3853 |
| ами/ями | | | | | Ins | | | 22971,8488 |
| ах/ях | | | | | | Loc | | 36574,8502 |

# Номинатив, DD

2. Для каждого глагола запоминаем, какие формы употреблялись в этой позиции
3. Смотрим, каково число глаголов, у которых количество попаданий в каждую «правильную» клетку больше, чем в любую другую

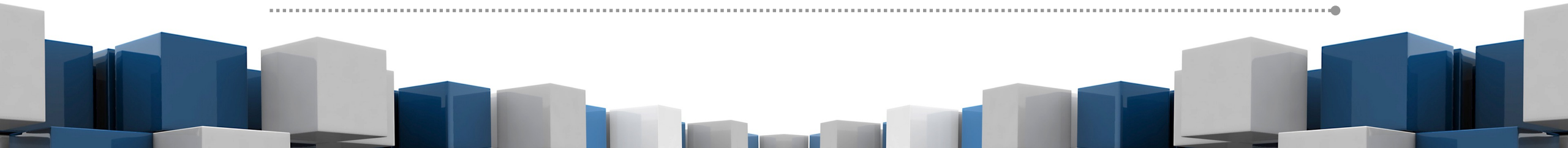


# Номинатив, DD

2. Для каждого глагола запоминаем, какие формы употреблялись в этой позиции
3. Смотрим, каково число глаголов, у которых количество попаданий в каждую «правильную» клетку больше, чем в любую другую:

0.42 («крайнее левое имя»)

0.40 («ближайшее слева имя»)

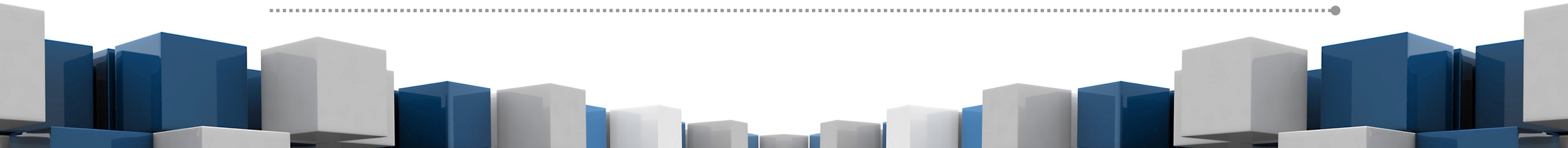


# Аккузатив, DD & UG

1. Смотрим на второго участника, дополнительного к номинативному у финитных форм или на единственного участника нефинитных форм
2. Случайно с вероятностью ½ решаем, является ли глагол переходным
3. Случайно выбираем на роль аккузативного существительное из оставшихся кандидатов

# Аккузатив, UG

4. Смотрим, какие клетки оказались наиболее частотными для всех встретившихся в корпусе глаголов



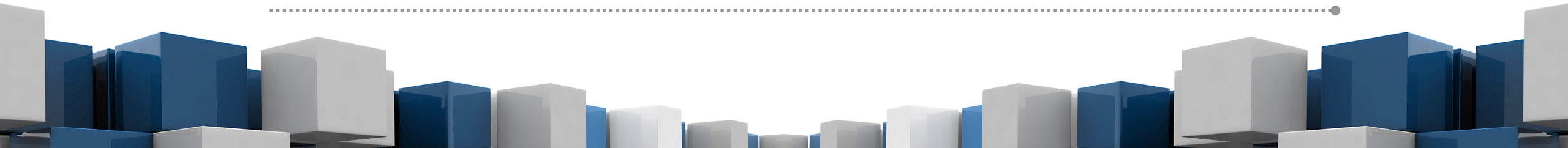
# Аккузатив, UG

4. Смотрим, какие клетки оказались наиболее частотными для всех встретившихся в корпусе глаголов

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| а/я | Nom | Acc | Gen | | | | | 26834 |
| с | Nom | Acc | Gen | | | | | 26588 |
| о | Nom | Acc | | | | | | 29155 |
| е | Nom | Acc | | Dat | | Loc | | 28879 |
| ы | Nom | Acc | Gen | | | | | 30457 |
| и | Nom | Acc | Gen | Dat | | Loc | | 29071 |
| у/ю | | Acc | | Dat | | Loc | GenP | 35148 |
| ой | | | | | Ins | | | 33316 |
| ом/ем | | | | | Ins | | | 34248 |
| ей | | Acc | Gen | | Ins | | | 30400 |
| ов | | Acc | Gen | | | | | 30168 |
| ам/ям | | | | Dat | | | | 32611 |
| ами/ями | | | | | Ins | | | 40288 |
| ах/ях | | | | | | Loc | | 34383 |

# Аккузатив, DD

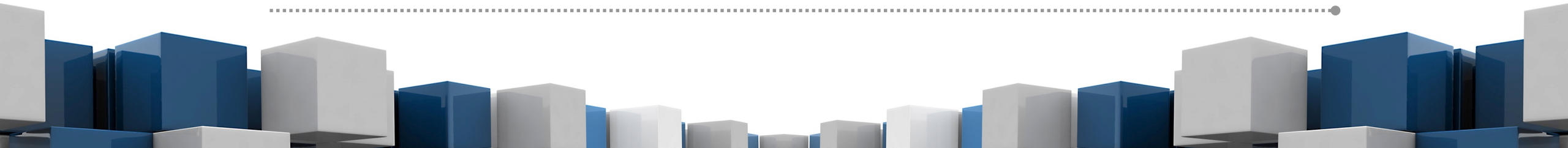
4. Смотрим, каково число глаголов, у которых количество попаданий в каждую «правильную» клетку больше, чем в любую другую



# Аккузатив, DD

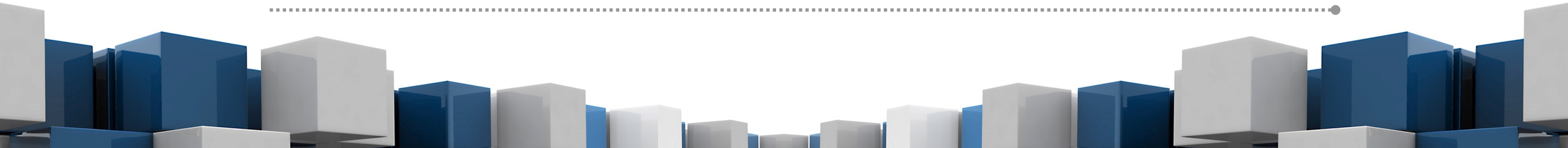
4. Смотрим, каково число глаголов, у которых количество попаданий в каждую «правильную» клетку больше, чем в любую другую

```
% глаголов с правильными падежами =  52.663
```



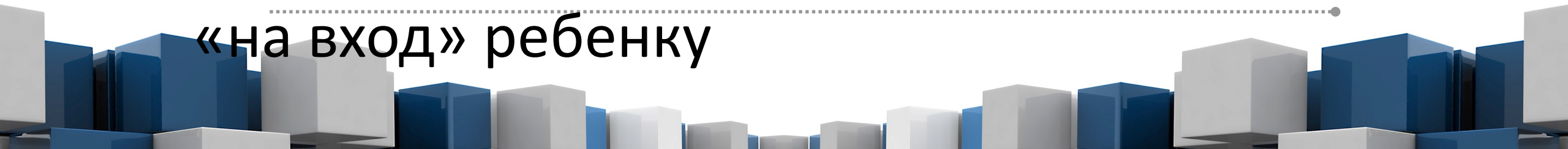
# Выводы

1. Предложный падеж демонстрирует хорошую подверженность DD-подходу к усвоению несмотря на небольшое к-во словоформ
2. Номинатив может быть уверенно усвоен при подходе UG и демонстрирует плохие результаты при DD



# На будущее

- Доработать дизайн эксперимента с аккузативом
- «Смешанный» (DD, then UG) подход, см. (Radford 1990) a.o.: функциональная структура усваивается после лексической
- Объем СинТагРуса: навскидку примерно на порядок меньше, чем может быть данных «на вход» ребенку



# Спасибо
# за внимание!