Представление устойчивых лексических сочетаний в компьютерном тезаурусе RussNet

Азарова Ирина Владимировна

Синопальникова Анна Алексеевна

Кафедра математической лингвистики СПбГУ

Смрж Павел

Технологический университет Брно (Чехия)

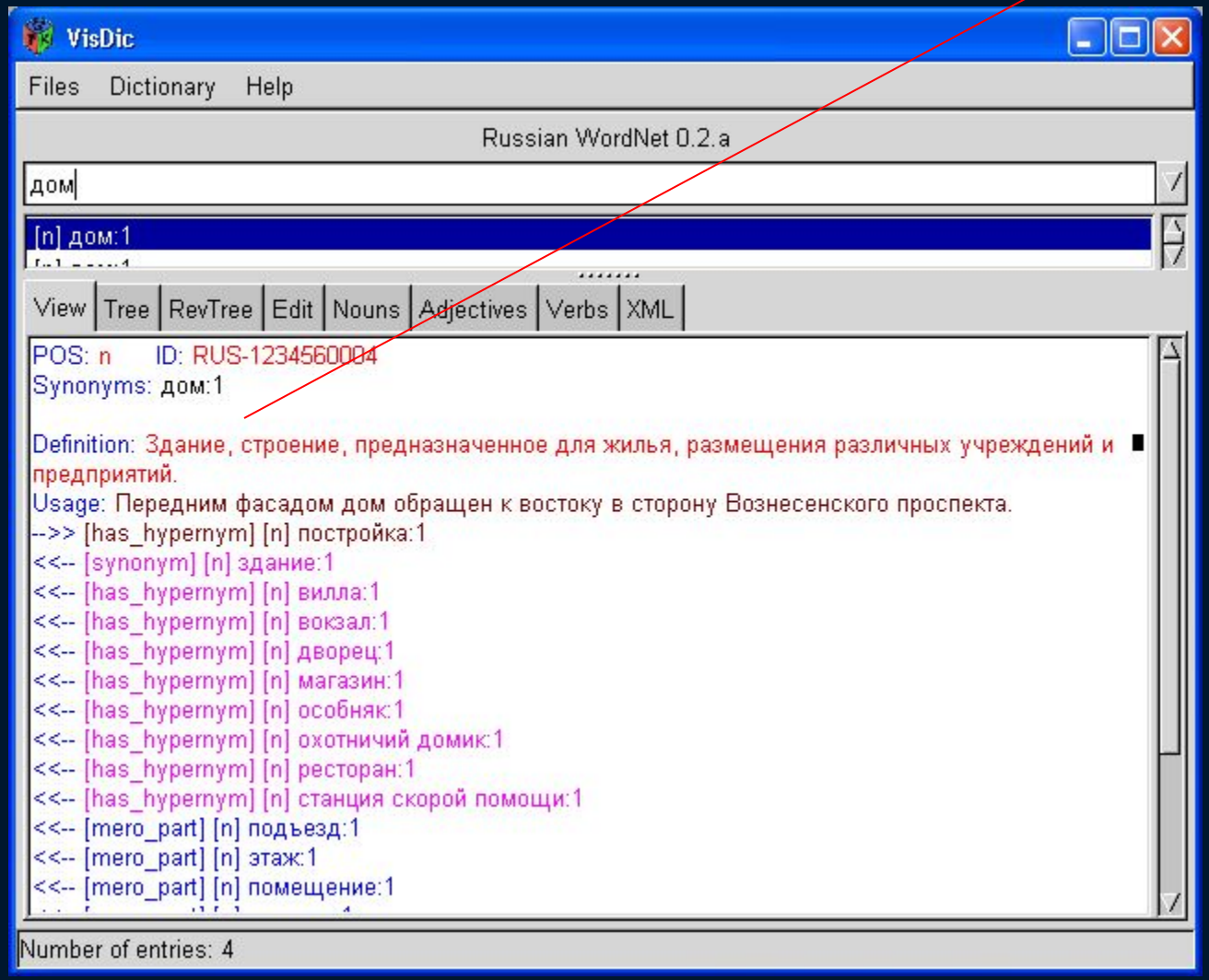
Компьютерный тезаурус RussNet

Является моделью человеческого лексикона:

- **в тезаурус включены смыслы, которые имеют словесное выражение в русском языке**
- **элементарной единицей является синонимический ряд (СИНСЕТ)**
- **между синсетами установлены семантические отношения**
- **семантическое отношение род-вид организует синсеты в набор деревьев**

Пример словарной статьи RussNet

Синсет, состоящий из одного слова «дом»



Корпус современных текстов Бокренок 3

Объем: 21 млн. словоупотреблений

Временной интервал: 1985 –2004

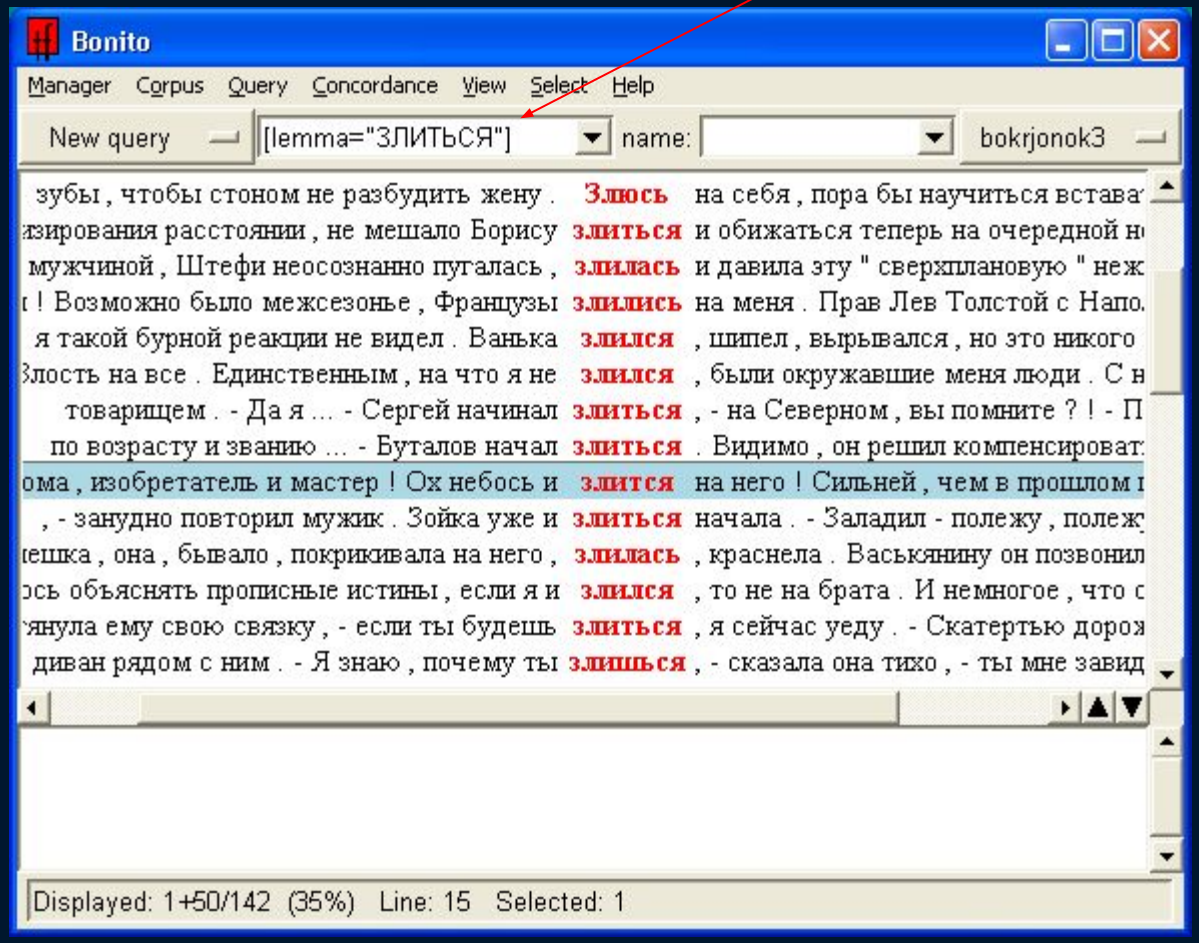
Типы текстов:

40% – газетные тексты
30% – научно-популярные тексты
20% – фрагменты художественных текстов
10% – тексты законов, договоров…

- баланс стереотипности, информативности и экспрессивности текстов
- отсутствие больших фрагментов текста

Отбор контекстов в программе Бонито

Запрос на поиск контекстов слова



Контекстные маркеры значений лексем в текстах

Контекстные маркеры можно задавать в виде

- **рамок валентности – устойчивых семантических и морфо-синтаксических признаков контекстов;**
- **сочетаний слов, имеющих собственное связанное значение и форму выражения.**

Схема описания активной валентной рамки

V_n <N, ROLE, SEM, SYN, f_o>…

N – порядковый номер (при объективном порядке слов)

ROLE семантическая роль в пропозиции;
значения: СУБЪЕКТ, ОБЪЕКТ1, ОБЪЕКТ2…
SEM семантический тип слов, заполняющих позицию;
значения: ЧЕЛОВЕК, ПРЕДМЕТ, АРТЕФАКТ…
SYN частотная форма морфо-синтаксичского заполнения позиции; ***значения:*** acc, dat, abl, «на» + acc …
F_o 1 – обязательная (> 66-85%),
0 – факультативная (> 15-33%)

Валентная рамка RussNet

Направиться (358 контекстов - 100%)

“двинуться в каком-л направлении”

Рамка валентностей (2 обязательные позиции)

(1) ***субъект движения***

N_1 [лицо / группа лиц] СУБЪЕКТ

(2) ***направления движения***

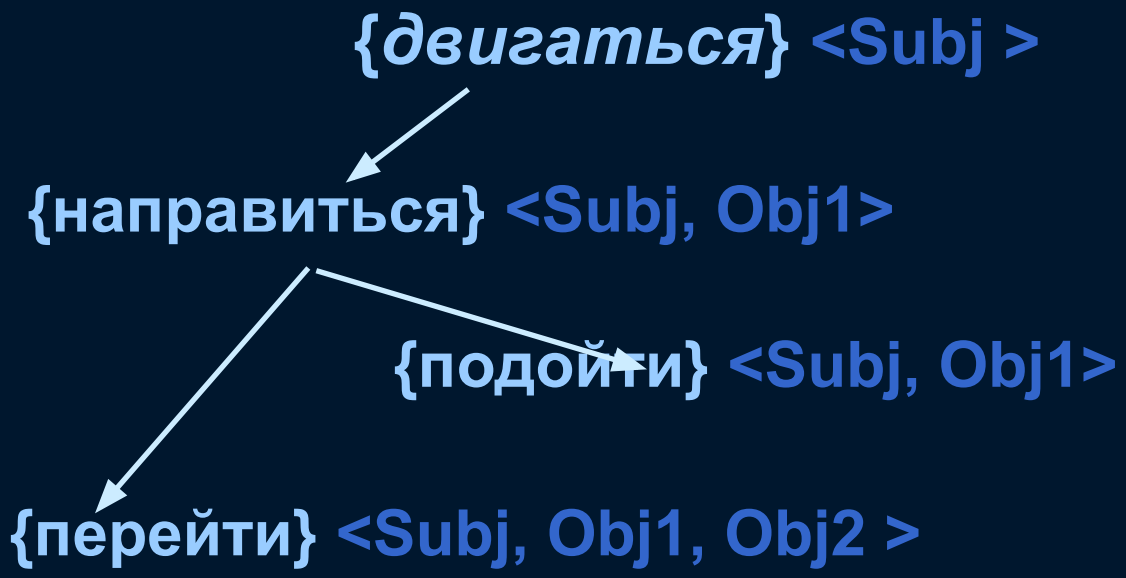
"к + N_3" (44%) [локализация/ лицо] направление движения?

"в + N_4" (27%) [локализация] ОБЪЕКТ1

"в стороону + N_4" (4%)…

(*к дивану, к другу, к спуску, к нему…; в комнату, в деревню, в угол гостиной*)

Семантические роли валентных позиций в RussNet



- семантические деревья RussNet обладают конечным набором семантических ролей
- типы объектов в дереве устойчивы

Пассивная валентная рамка RussNet

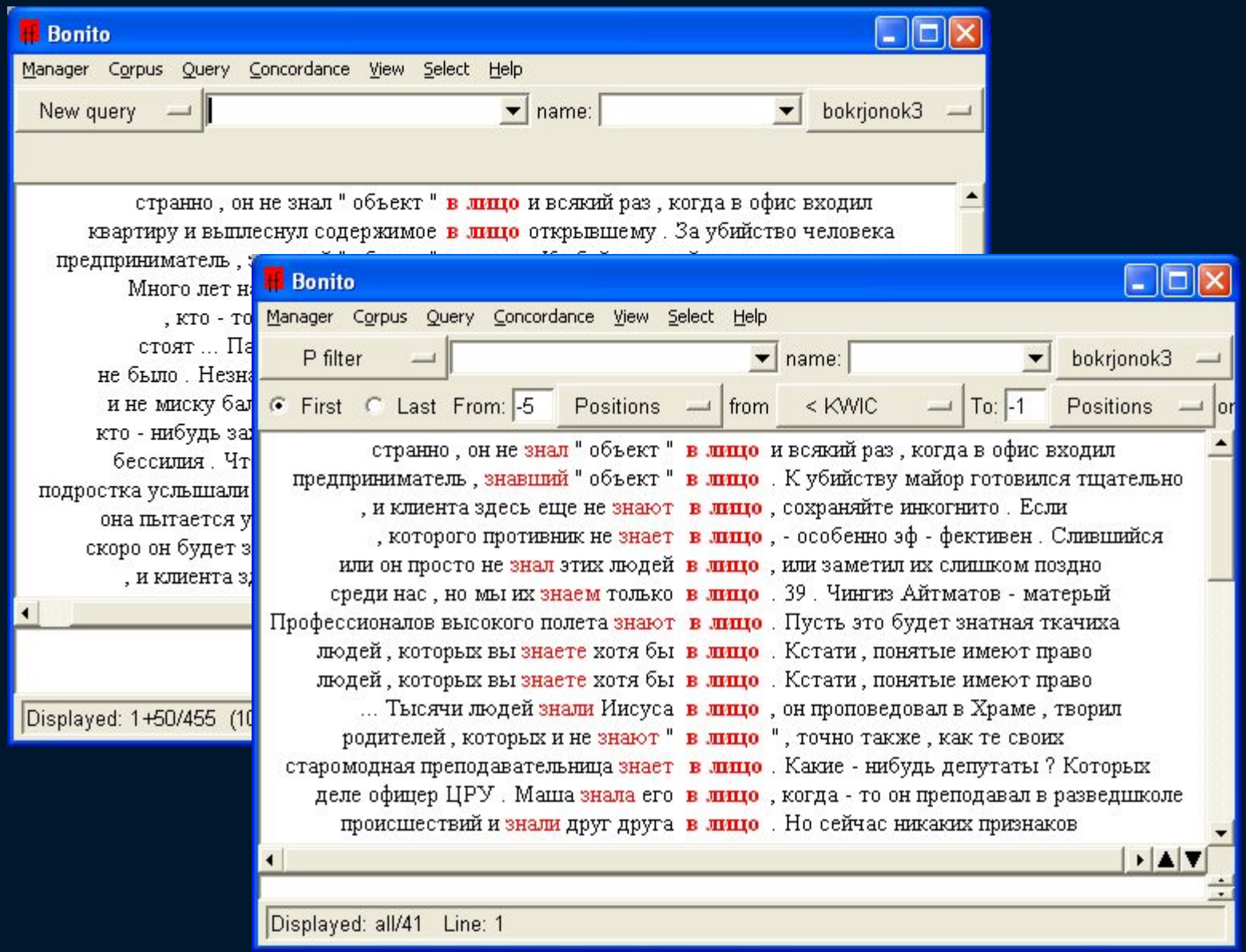
_V (SYN) <ROLE, SEM>…

SYN морфо-синтаксичская форма слова, имеющего пассивную валентность
в лицо "в + N_4"

ROLE семантическая роль в пропозиции;

SEM семантический тип слов, присоединяющих валентность
{сказать}, {знать}, {плюнуть}

Пассивная валентная рамка RussNet (2)



Словосочетания в wordnet-словарях

Принстонский WordNet

give up

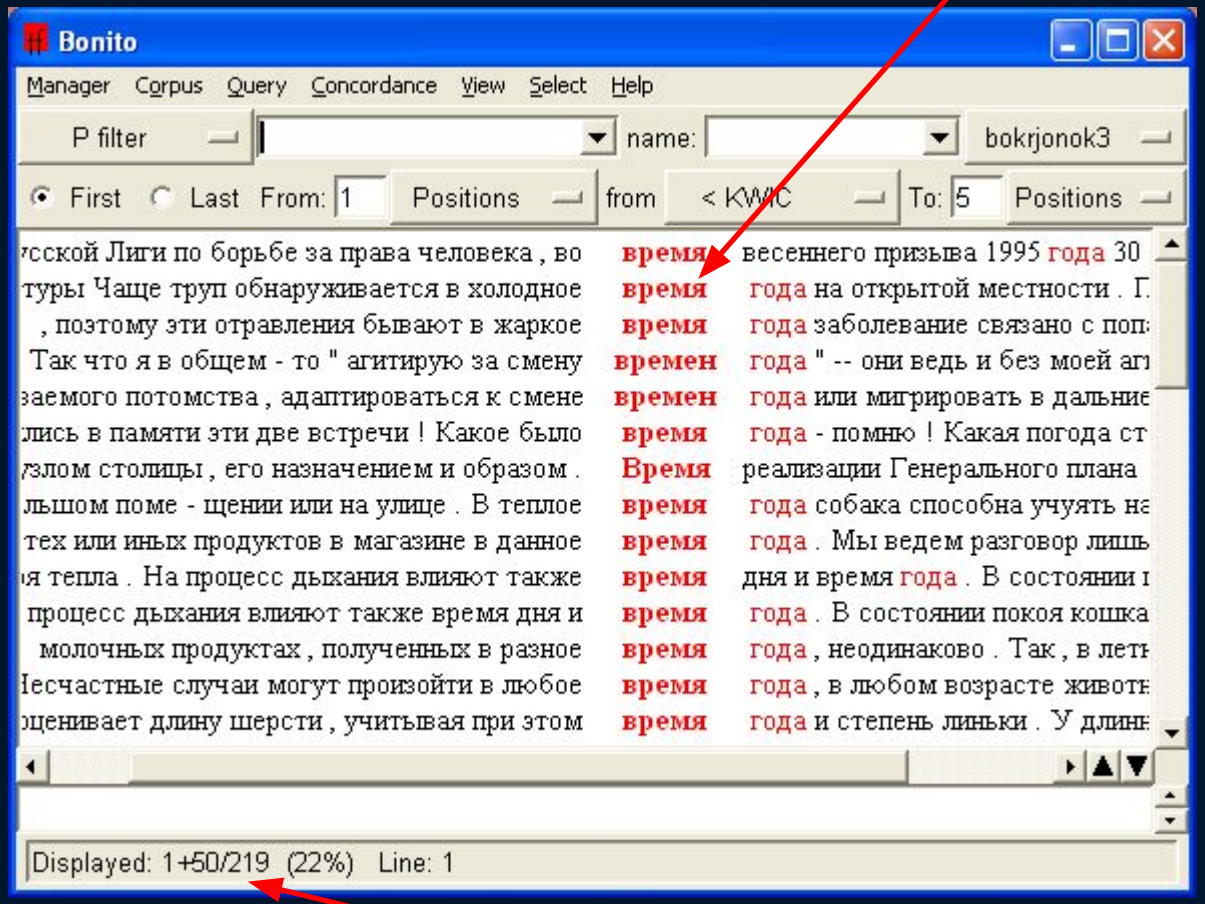
Sense 1
forfeit, give up, throw overboard, render, waive, forgo
=> abandon -- ("We abandoned the old car in the empty parking lot")

Sense 2
drop out, give up, throw in, throw in the towel, quit, leave, admit defeat, chuck up the sponge -- (give up in the face of defeat of lacking hope)

Sense 3
spare, give up, part with, dispense with
=> give -- (transfer possession of something concrete or abstract to somebody; "I gave her my money"; "can you give me lessons?")

Устойчивые словосочетания (1)

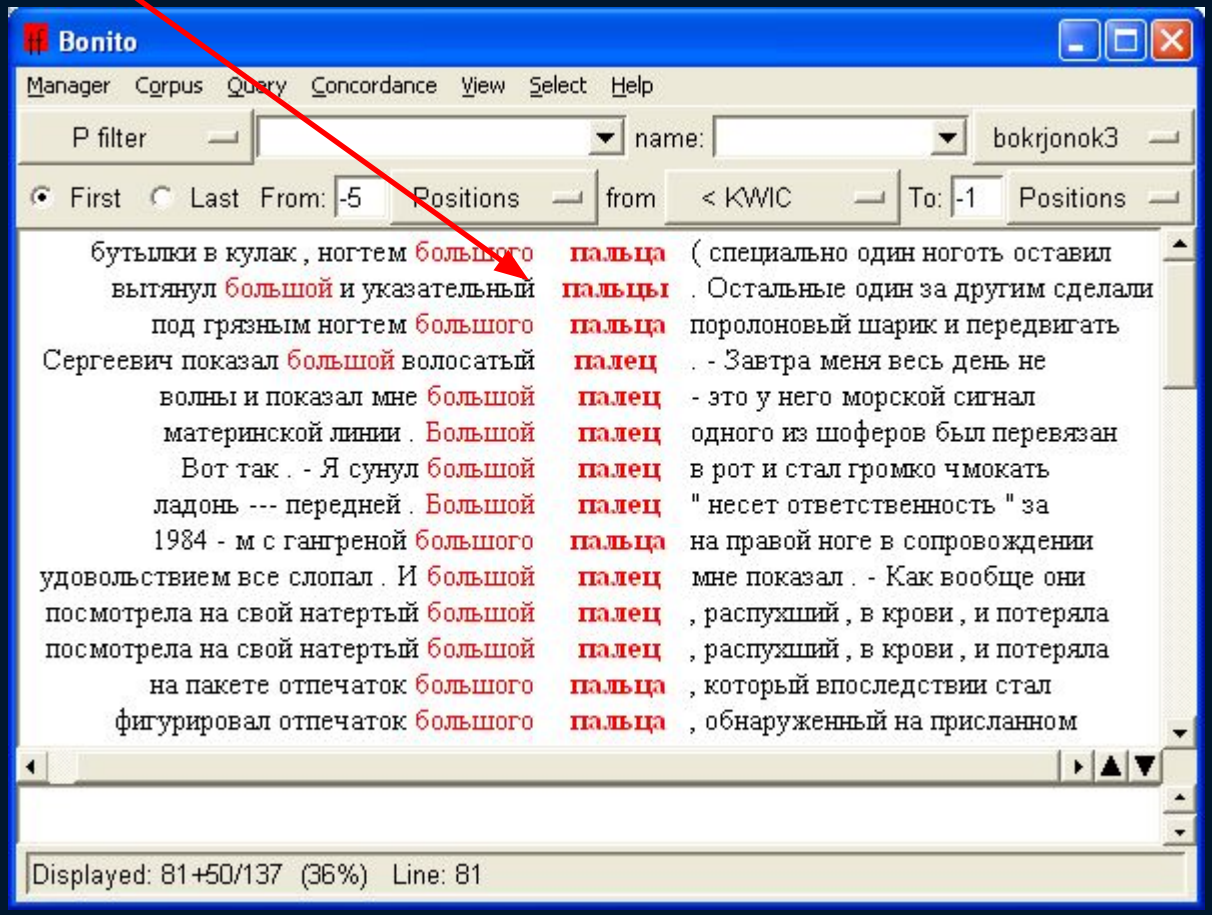
Контактность существительных



Воспроизводимость

Устойчивые словосочетания (2)

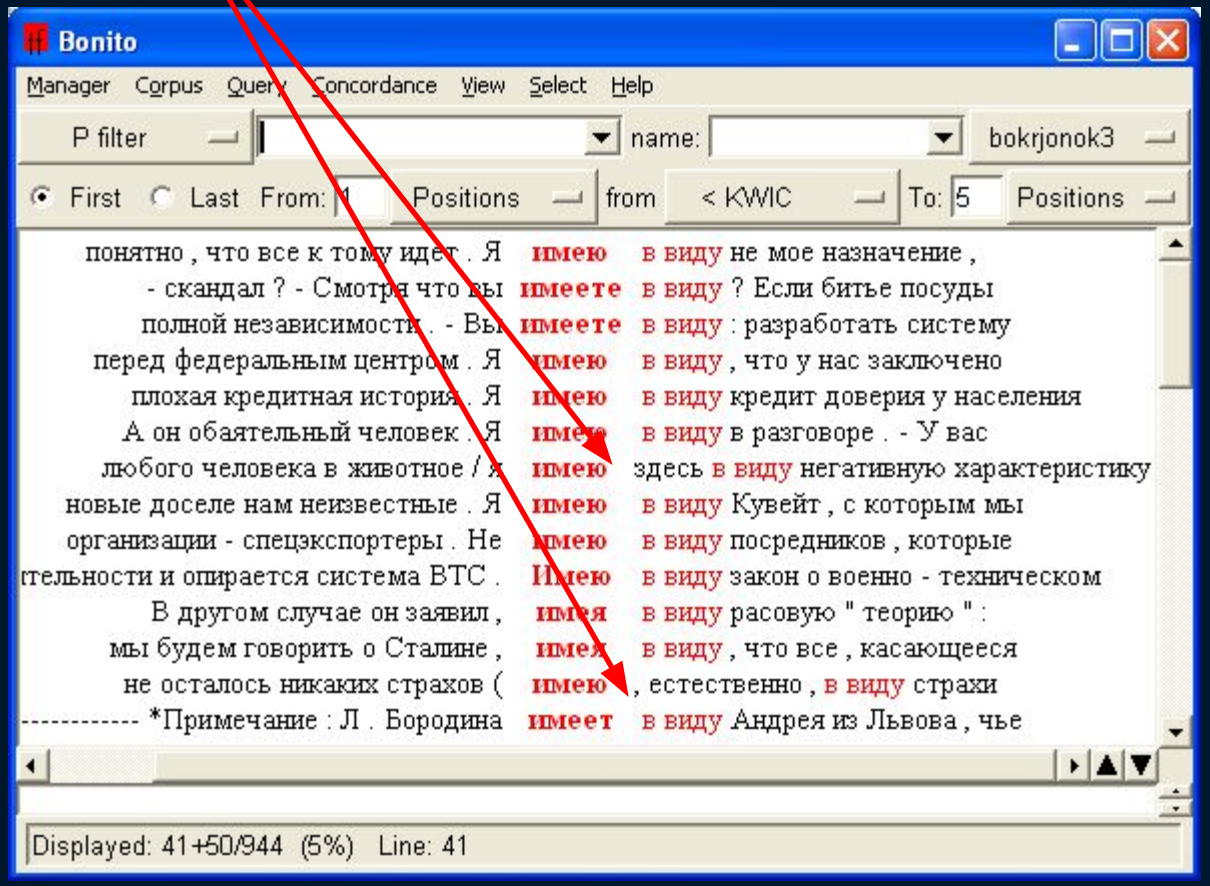
Контактность может нарушаться при перечислении.



Устойчивые словосочетания (3)

Глагольные словосочетания

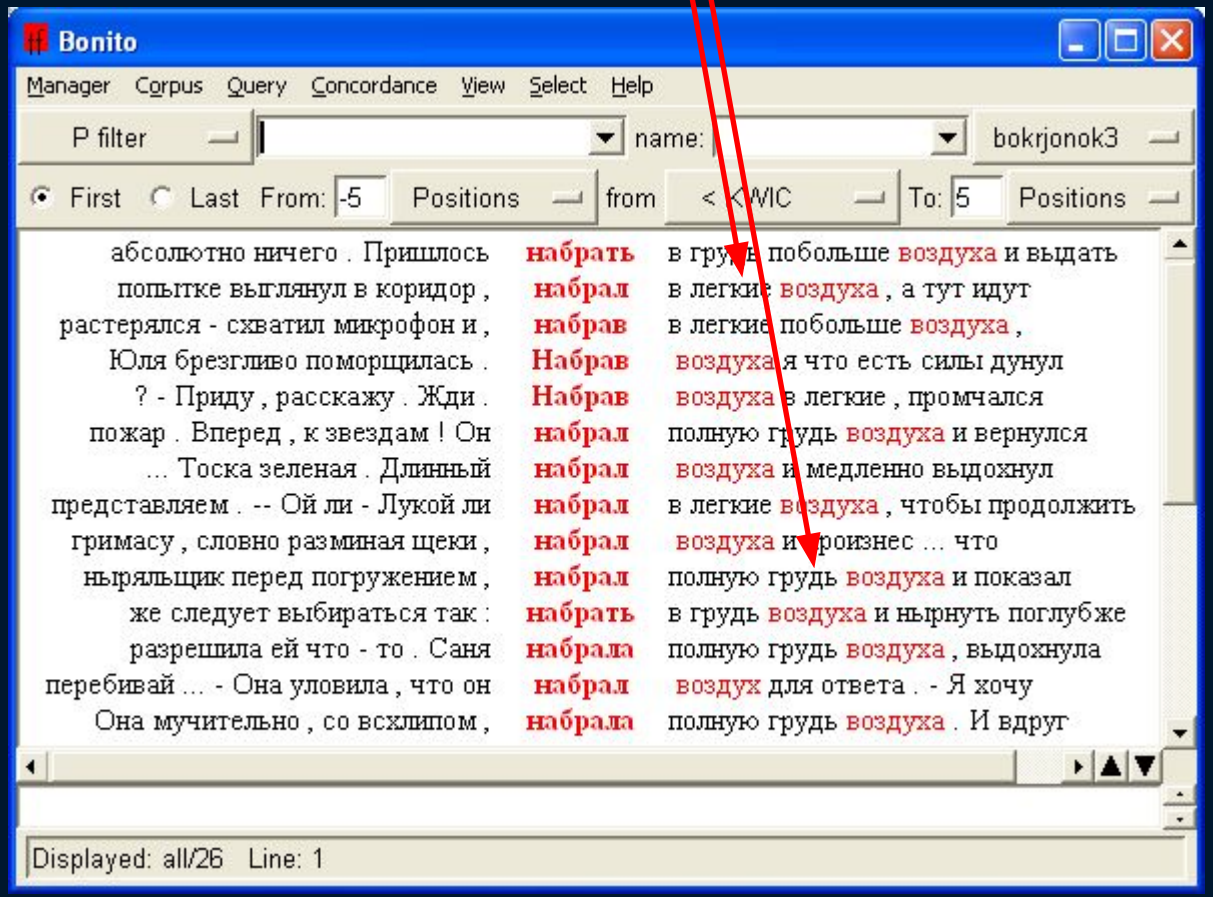
Контактность нарушается наречиями, частицами и проч.



Устойчивые словосочетания (4)

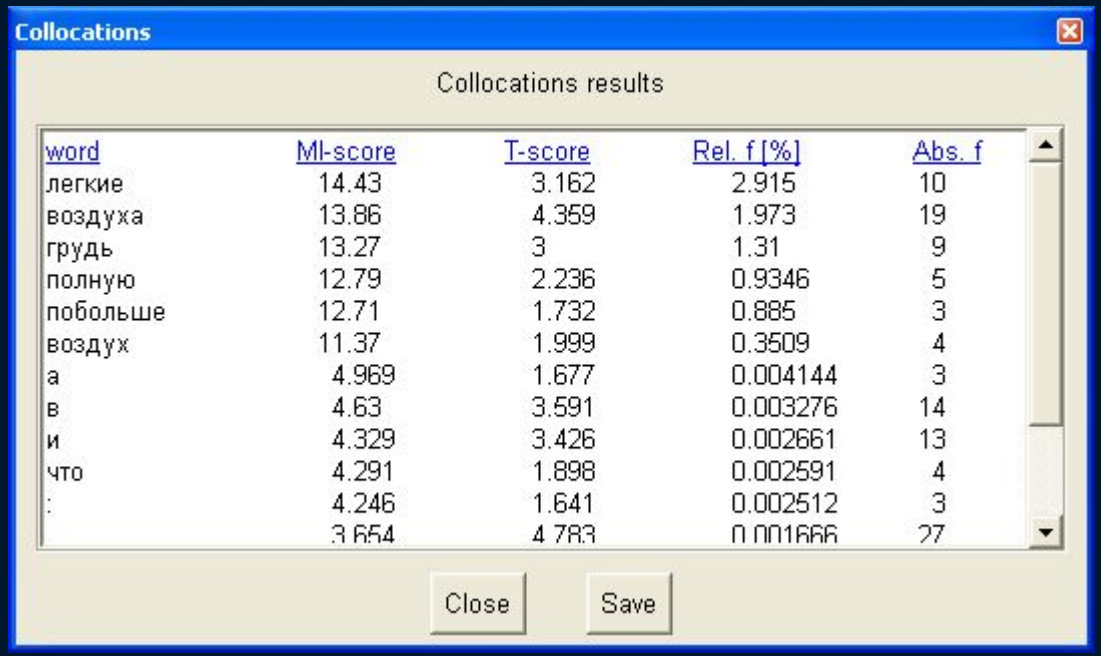
Глагольные словосочетания

В менее очевидных случаях контактность нарушается другими валентными позициями.



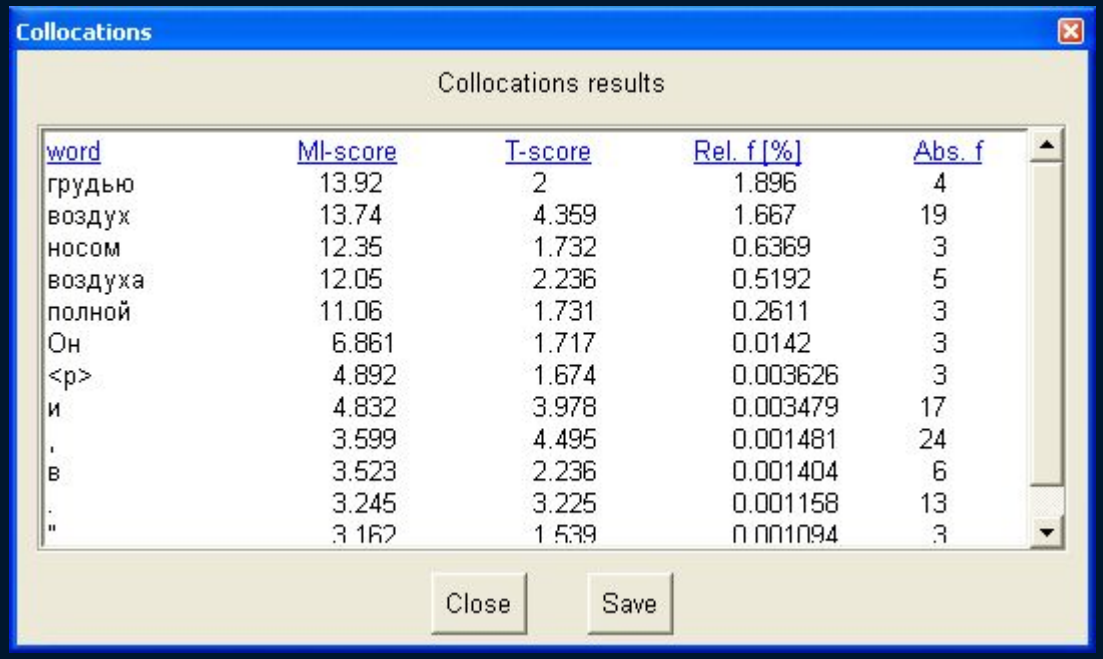
Устойчивые словосочетания (5)

Статистическая оценка «устойчивости» сочетания *набрать воздух(а)* в диапазоне ±5 слов. Коэффициент «неслучайности» MI показывает, насколько «информативно» данное сочетание слов. Доля контекстов с данным значением 3%, общее число вхождений глагола 664 (31 ipm)



Устойчивые словосочетания (6)

Статистическая оценка сочетаний *вдохнуть воздух* в диапазоне ±5 слов (частотность в корпусе – 5 ipm, 101 вхождение)



Контактная информация

RussNet

http://www.phil.pu.ru/depts/12/RN/