Количественные методы в филологии

Основные области применения количественных методов

- Корпусная лингвистика
- Сравнительно-историческое языкознание
- Социолингвистика
- Дискурс-анализ
- Стилистика (анализ лексики как основание для отнесения текста к определённому стилю)
- Атрибуция текстов
- Анализ поэтического текста

Контент-анализ как количественный метод анализа текстов

- Сущность контент-анализа заключается в том, чтобы **по внешним — количественным — характеристикам текста** на уровне слов и словосочетаний сделать **правдоподобные предположения о его плане содержания** и, как следствие, сделать выводы об особенностях мышления и сознания автора текста — его намерениях, установках, желаниях, ценностных ориентациях и т. д.

Контент-анализ как количественный метод анализа текстов

- Важнейшей категорией контент-анализа является **концептуальная переменная** — понятие, которое стоит в центре проводимого исследования («СВОЙ-ЧУЖОЙ», «ДЕМОКРАТИЯ», «ПРАВА ЧЕЛОВЕКА», «ЖЕНСКИЙ ВОПРОС»)
- В конкретном тексте концептуальная переменная представлена своими значениями — **языковыми представителями**.
- Так, концептуальная категория «СВОЙ-ЧУЖОЙ» в текстах может иметь следующие значения: *мой, наш, мы, я, привычный, знакомый, близкий vs. их, его, ее, он, она, оно, они, их, ее, его, непривычный, дальний, незнакомый.*

Этапы подготовки и проведения исследования

1) выбор материала—корпуса языковых данных;
2) выбор концептуальной переменной и определение ее значений — языковых репрезентантов выбранного понятия в тексте;
3) выбор единицы кодирования; значения К-переменной могут приписываться текстам, их фрагментам, абзацам, предложениям и отдельным словам и словосочетаниям (например, заголовкам).

Этапы подготовки и проведения исследования

4) отбор кодировщиков (выбор программ) и формулировка инструкций по кодированию; существует два вида контент-анализа — **жесткий и мягкий** (выявляются не только явные, но и неявные, имплицитные вхождения переменной);

5) кодировка данных;

6) подсчет данных и интерпретация результатов.

Проблема семантической достоверности

- Необходимо учитывать многозначность языковых выражений, являющихся значениями К-переменной.

*После этого тихо тлевшая **война** перешла в открытые боевые действия. «Мослифт» полностью перестал обращаться на тот самый завод, чьи технологии — капельная пропитка статоров, централизованная нарезка канатов с обваркой концов, автоматизированная очистка редукторов главного привода и тому подобные лифтовые премудрости, — существенно улучшают качество ремонта (анализ К-переменной «ВОЙНА-МИР»).*

Исследование политических метафор с помощью контент-анализа

- Х. де Ландшер на материале голландского политического дискурса за период с 1831 по 1981 гг. [Christ'l de Landtsheer 1991] попытался установить возможные корреляции между частотой использования в политическом дискурсе политических метафор и периодами политико-экономических кризисов

Степень метафоричности дискурса

Степень метафоричности для разных метафор может различаться, для общей оценки силы метафоры была введена переменная референциальной интенсивности I, которая вычислялась по следующей формуле:

$$I = \frac{1w + 2n + 3s}{t},$$

- w — количество «слабых» метафор (реализация стандартных метафорических переносов значения)
- n — обычные конвенциональные метафоры, не фиксированные как словарные значения
- s — абсолютно новые, креативные метафоры;
- t — общее количество метафор.

Тип метафорической модели

- Сила воздействия метафоры связана не только с новизной или общепринятостью, но и с типом самой метафорической модели. Понятно, что метафорические модели ВОЙНЫ, СМЕРТИ, БОЛЕЗНИ более конфликтны, чем метафорические модели СТРОЕНИЯ, ПУТИ. Для отражения конфликтности была введена еще одна переменная — переменная содержания D, которая вычислялась по следующей формуле:

$$D = \frac{1p + 2n + 3po + 4d + 5sp + 6m}{t},$$

- p — стертые метафоры;
- n — метафоры природы;
- po — политические и интеллектуальные метафоры;
- d — метафоры смерти и бедствий;
- sp — игровые и спортивные метафоры;
- m — метафоры болезни;
- t — общее количество метафор.

Результаты исследования

- Результаты кодирования и вычисленные переменные интенсивности и содержания были сопоставлены с имеющимися в статистических справочниках данными по безработице и динамике оптовых цен.
- Оказалось, что динамика значений переменных I и D коррелирует с динамикой безработицы: **чем выше безработица, тем выше значение переменной интенсивности и переменной содержания**.
- Интересно, что оценка абсолютной частоты использования метафор в меньшей степени отражает степень корреляции, чем переменные I и D.

Семантический анализ текста

- Семантическая структура устного спонтанного текста: социолингвистическое варьирование (Павлова Д.С., Ерофеева Е.В.)
- В рамках исследования использовался **метод графосемантического моделирования текста** (разработан К. И. Белоусовым, Н.Л. Зелянской, Д.А. Барановым) с применением информационной системы «Семограф» (http://semograph.com)

Графосемантическое моделирование

- «Графосемантическое моделирование представляет собой метод графической экспликации структурных связей между семантическими компонентами одного множества» [Белоусов 2009: 31].
- Таким множеством может являться любой текст.

Алгоритм графосемантического моделирования

1) проведение компонентного анализа отобранного материала, т. е. выделение контекстов и компонентов из всего массива материала, в нашем случае – из текста;

2) проведение полевого анализа выделенных компонентов, который подразумевает объединение компонентов в поля (классы);

3) генерация семантической карты, отражающей связи между полями в пределах всей выборки;

4) графическая экспликация результатов анализа в виде графа;

5) интерпретация полученной модели.

Семантический анализ спонтанного монолога

- **Контекст**—синтагма (минимальное интонационное и синтаксическое целое);
- **Компонент** –знаменательное слово (исключаются дискурсивные слова, заполнители пауз хезитации и пр.)
- **Поле**=семантическое поле=микротема;

Алгоритм анализа

1) **Построение семантической классификации компонентов** (один компонент может относиться к разным группам; например, слово университет, которое встречается в контексте *я филолог закончила университет...* Относится одновременно к микротемам ОБРАЗОВАНИЕ и МЕСТО (поскольку информант в данном случае говорит именно о месте, где она получала образование).

2) Построение семантической карты (выявление связей между всеми полями текста)

3) Построение графа

Состав микротемы «я»

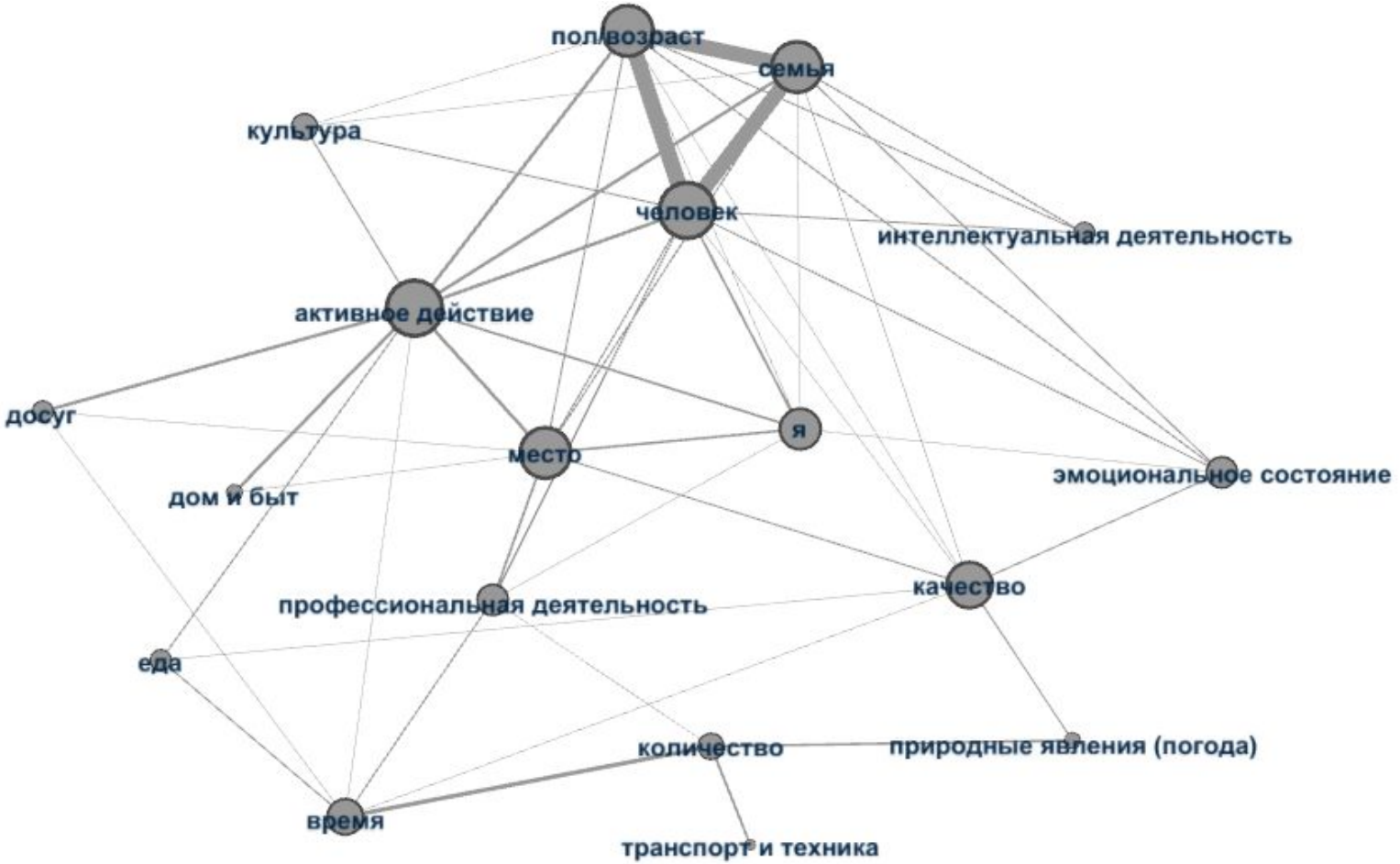
- Местоимение *я* и его формы (*мне, меня, мной* и др.),
- местоимение *мы* и его формы (*нам, нами, нас* и др.);
- глаголы в форме 1 лица ед.ч. (*помню, люблю, имею, считаю, работаю* и др.);
- глаголы в форме прошедшего времени (*сходил, стоял, считал, поступил, учился* и др.);
- глаголы в форме множественного числа, подразумевающие действия, которые совершает сам рассказчик вместе с родственниками /друзьями /коллегами (*осматривали, заняли, ездим, проживаем* и др.);
- собственные имена самих информантов (*Александр, Лариса, Лида Нечаева, Чуклинов Антон Викторович, Наталья, Тамара Леонидовна*).

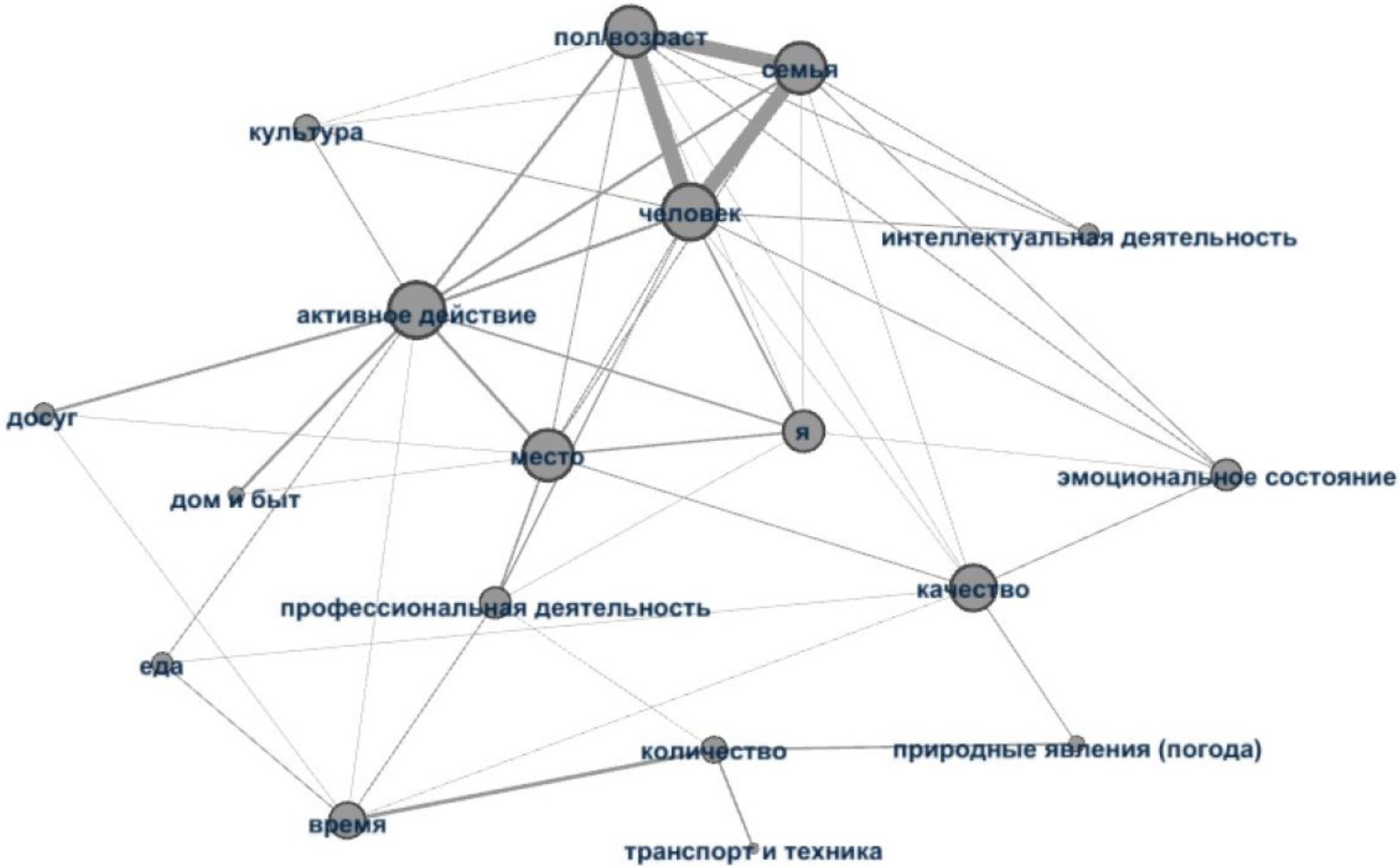
Микротема	Частота	
	абс.	%
ВРЕМЯ	22	13,5
АКТИВНОЕ ДЕЙСТВИЕ	19	11,7
ПРОФЕССИОНАЛЬНАЯ ДЕЯТЕЛЬНОСТЬ	12	7,4
ЧЕЛОВЕК	10	6,1
КОЛИЧЕСТВО	9	5,5
СЕМЬЯ	8	4,9
МЕСТО	8	4,9
КАЧЕСТВО	7	4,3
ПОЛ/ВОЗРАСТ	7	4,3
КУЛЬТУРА	6	3,7
ПРИРОДНЫЕ ЯВЛЕНИЯ (ПОГОДА)	5	3,1
ИНТЕЛЛЕКТУАЛЬНАЯ ДЕЯТЕЛЬНОСТЬ	5	3,1
ТРАНСПОРТ И ТЕХНИКА	4	2,5
ЭМОЦИОНАЛЬНОЕ СОСТОЯНИЕ	3	1,8
Я	2	1,2
ДОСТИЖЕНИЕ ЦЕЛИ	2	1,2
ЧУВСТВЕННОЕ ВОСПРИЯТИЕ	2	1,2
ДОМ И БЫТ	2	1,2
ЕДА	1	0,6
ОЦЕНКА	1	0,6
СОСТОЯНИЕ	1	0,6

Сильные связи микротем: метод графосемантического моделирования (базовая методика классификации компонентов)

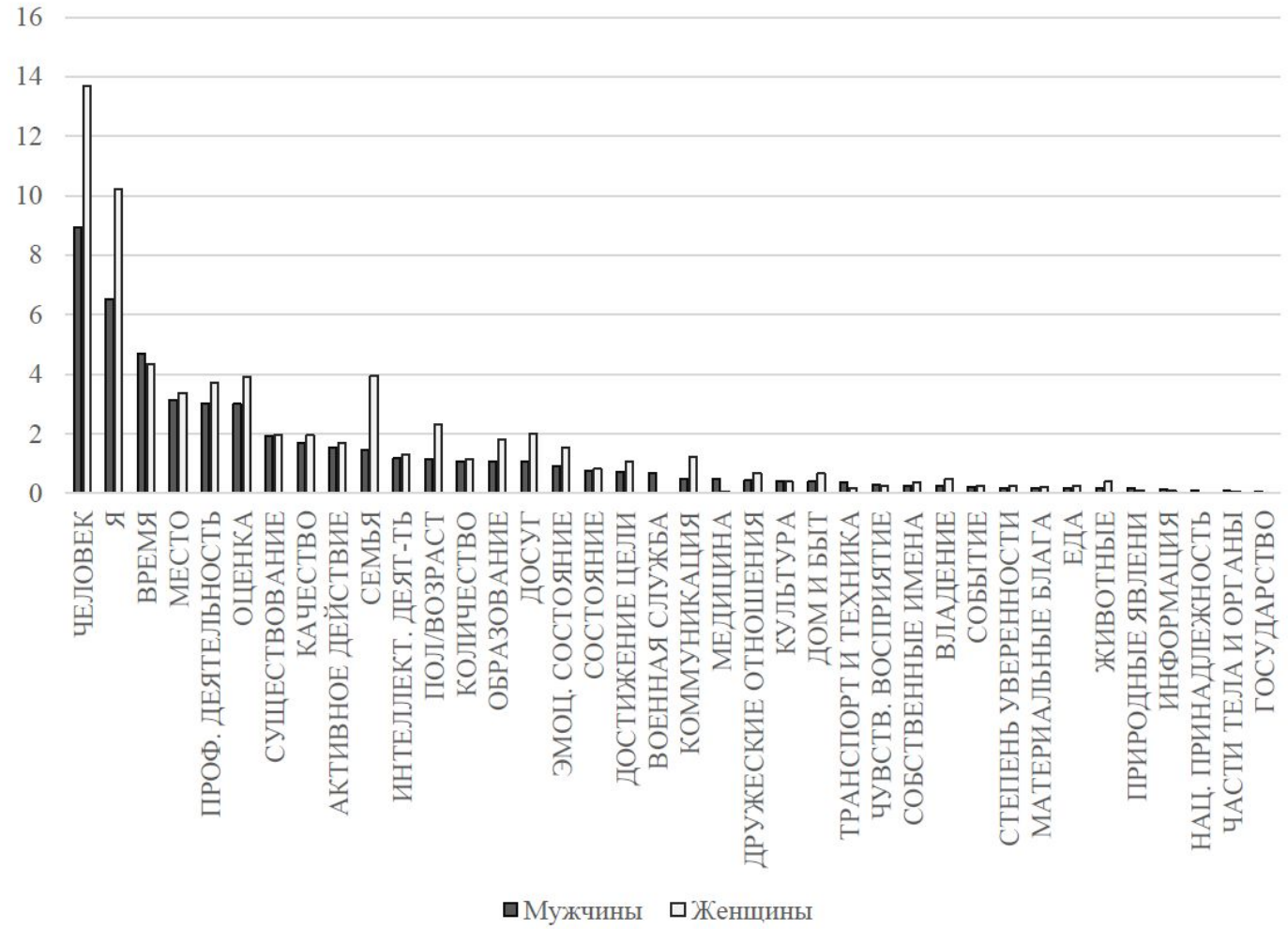
Микротемы	Сила связи	
	абс.	отн.
ЧЕЛОВЕК – ПОЛ/ВОЗРАСТ	7	1,00
ЧЕЛОВЕК – СЕМЬЯ	7	0,87
ВРЕМЯ – АКТИВНОЕ ДЕЙСТВИЕ	7	0,37
СЕМЬЯ – ПОЛ/ВОЗРАСТ	6	0,86
КОЛИЧЕСТВО – ВРЕМЯ	5	0,55
ВРЕМЯ – ЧЕЛОВЕК	5	0,50
ВРЕМЯ – ПОЛ/ВОЗРАСТ	4	0,57
ВРЕМЯ – СЕМЬЯ	4	0,50
ВРЕМЯ – МЕСТО	4	0,50
МЕСТО – АКТИВНОЕ ДЕЙСТВИЕ	4	0,50
ВРЕМЯ – ПРОФЕССИОНАЛЬНАЯ ДЕЯТЕЛЬНОСТЬ	4	0,33

Графосемантическое моделирование

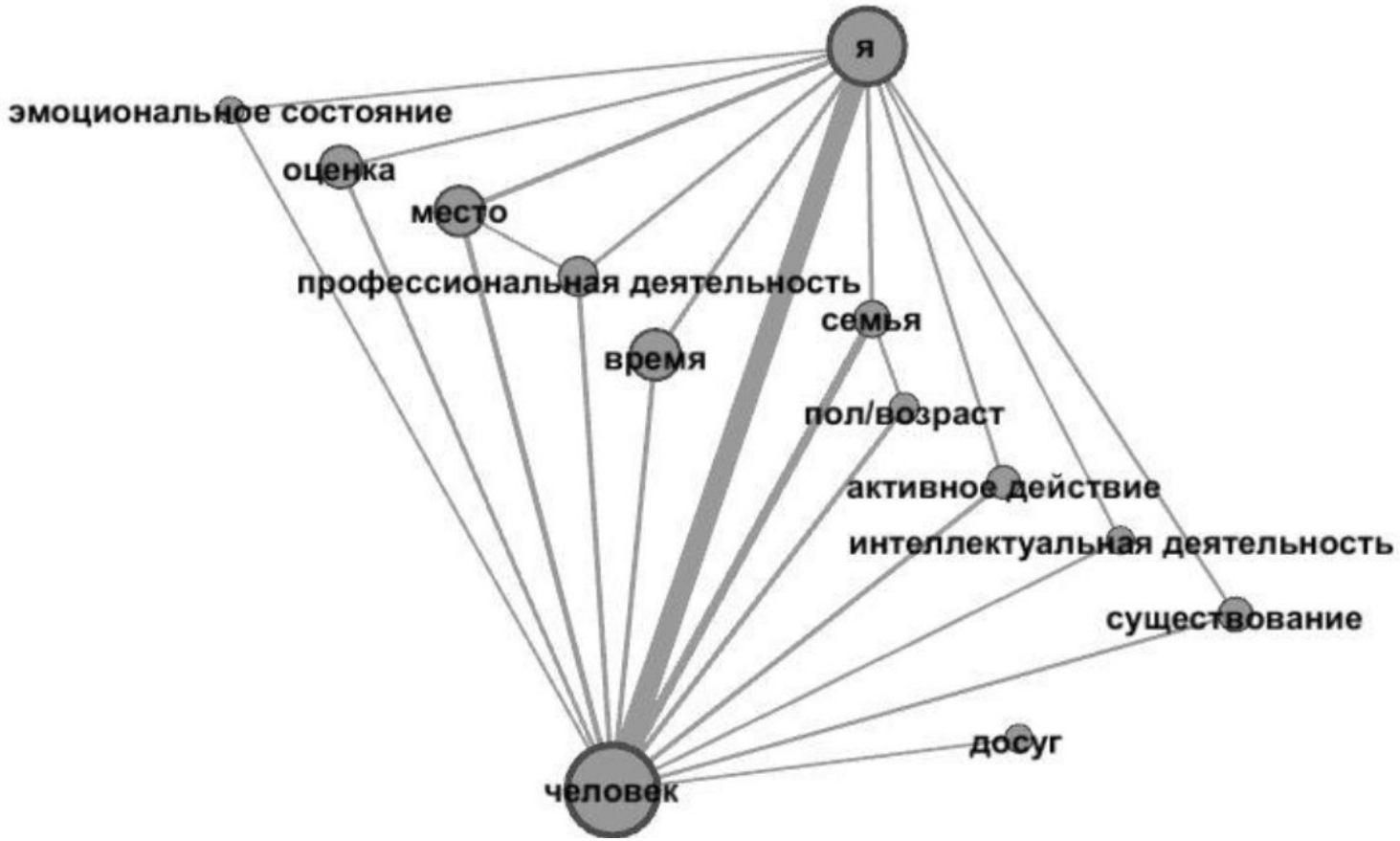



пол/возраст
семья
культура
человек
интеллектуальная деятельность
активное действие
досуг
я
место
эмоциональное состояние
дом и быт
качество
профессиональная деятельность
еда
количество
природные явления (погода)
время
транспорт и техника

Относительный объём микротем в текстах мужчин и женщин

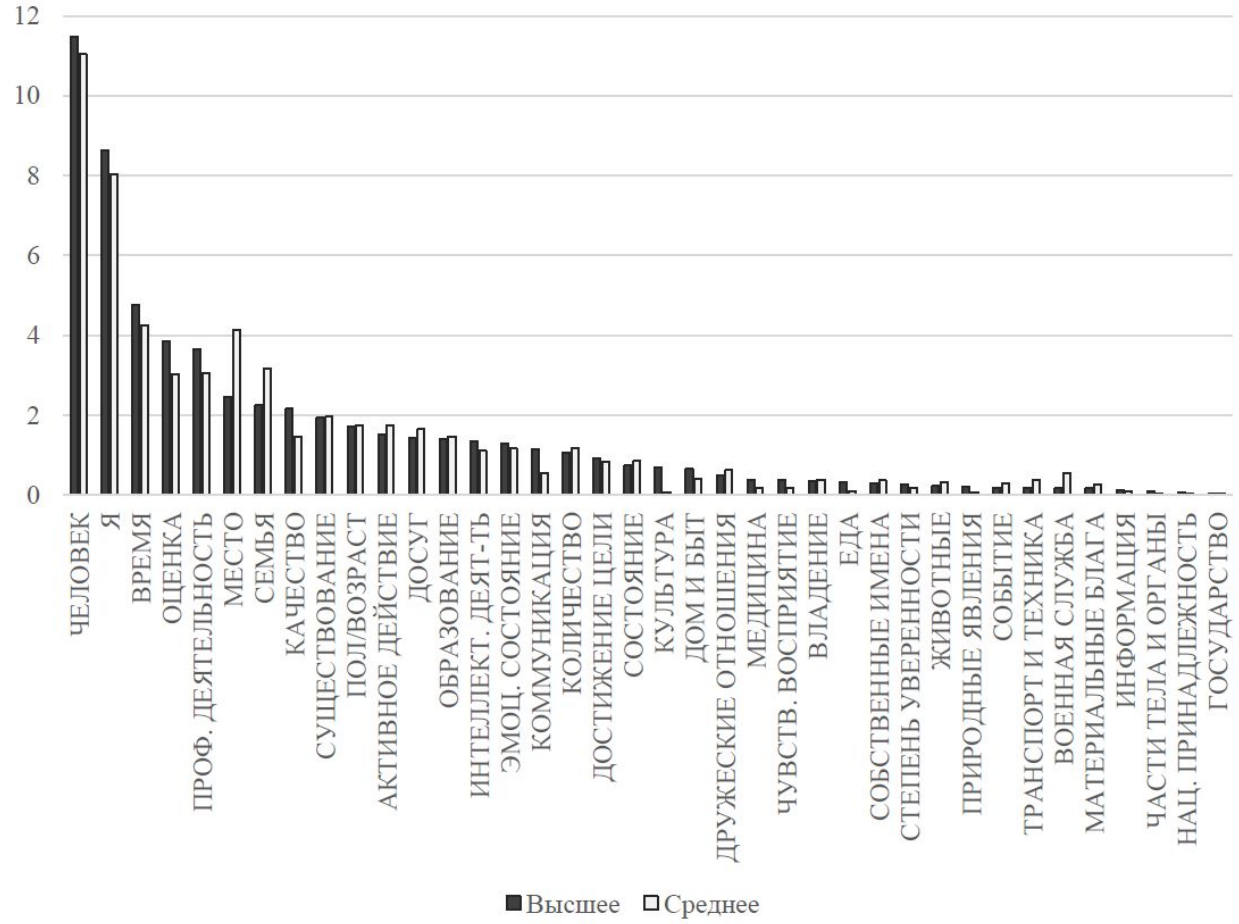


Модель семантической структуры текстов мужчин



Модель семантической структуры текстов женщин

Относительный объём микротем в текстах информантов с высшим и средним образованием



Автоматизированная обработка данных

- Специальные приложения для обработки данных и представления данных (система «Семограф» (http://semograph.com))
- Специализированные пакеты для обработки языковых данных есть в языках R и Python.
- Курс «R для лингвистов: программирование и анализ данных» (платформа «Открытое образование»)