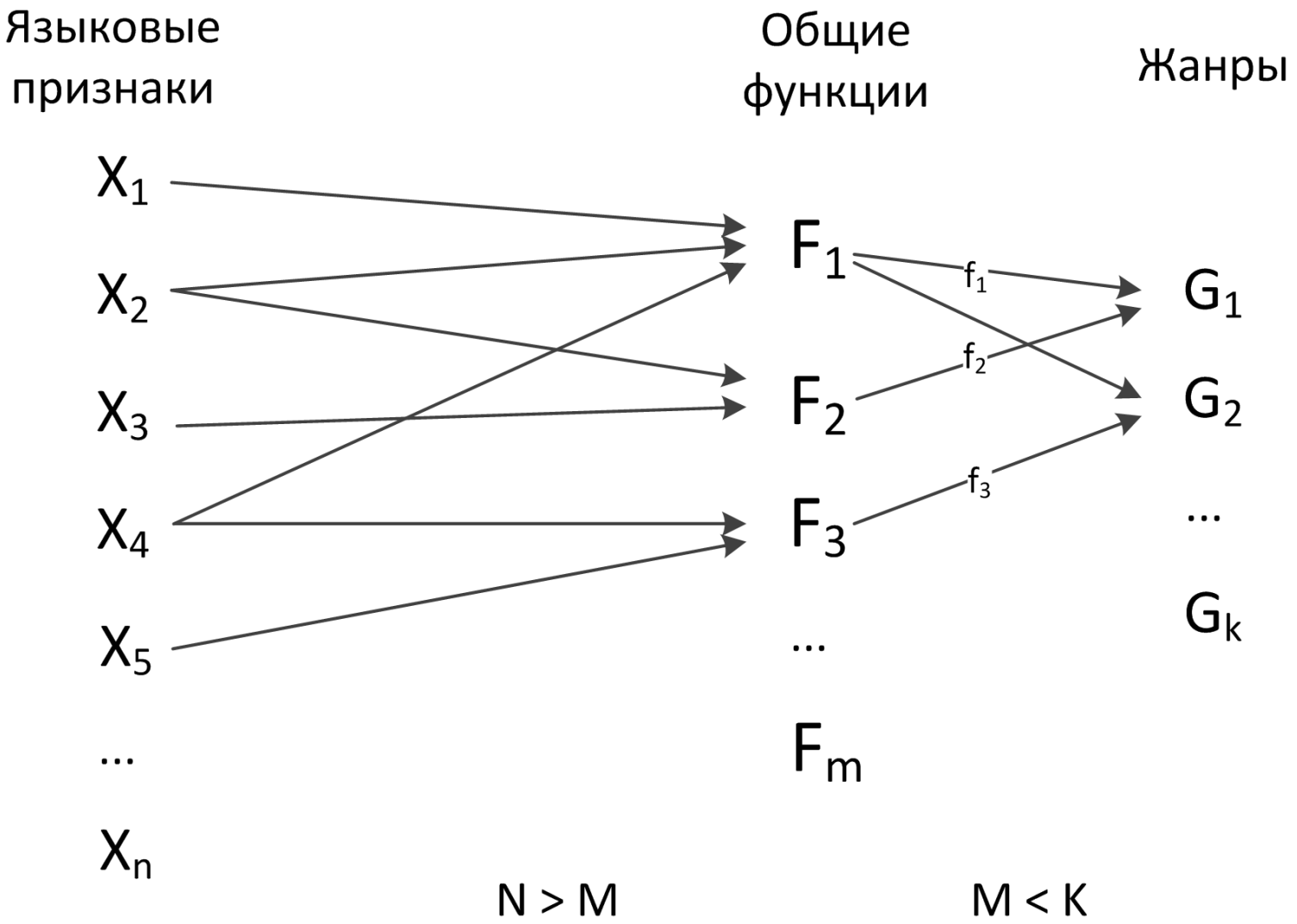
Применение Многомерного анализа к изучению языковой вариативности в русскоязычных Интернет-жанрах

Анисья Катинская

РГГУ

2 июня 2016 г.

Коммуникативно обусловленные вариации языковых признаков



Описание корпуса

Тексты	Слова	Предложения
774	815118	62582

Источники текстов

- Онлайн энциклопедии (wikipedia.org, krugosvet.ru, absurdpedia.net)
- Блоги (vk.com, livejournal.com, habrahabr.ru)
- Новости (ru.wikinews.org, chaskor.ru, ruposters.ru)
- Форумы (forum.hackersoft.ru, litforum.ru)
- Ресурсы юридических документов (base.garant.ru, consultant.ru)

Описание корпуса

Источники текстов

- Научные и научно-популярные порталы (sci-article.ru, cyberleninka.ru)
- Рекламные сайты(avito.ru, mvideo.ru, sportmaster.ru)
- Онлайн журналы (popmech.ru, afisha.ru)
- Онлайн библиотеки (lib.ru, wikisource.org),
- Политические платформы и платформы социальной журналистики (politonline.ru, politikus.ru, cont.ws)

Функциональные жанровые размерности

A1: argum В какой степени текст пытается убедить читателя поддержать какую-то точку зрения или позицию? (В большей степени, если аргументация очевидна).

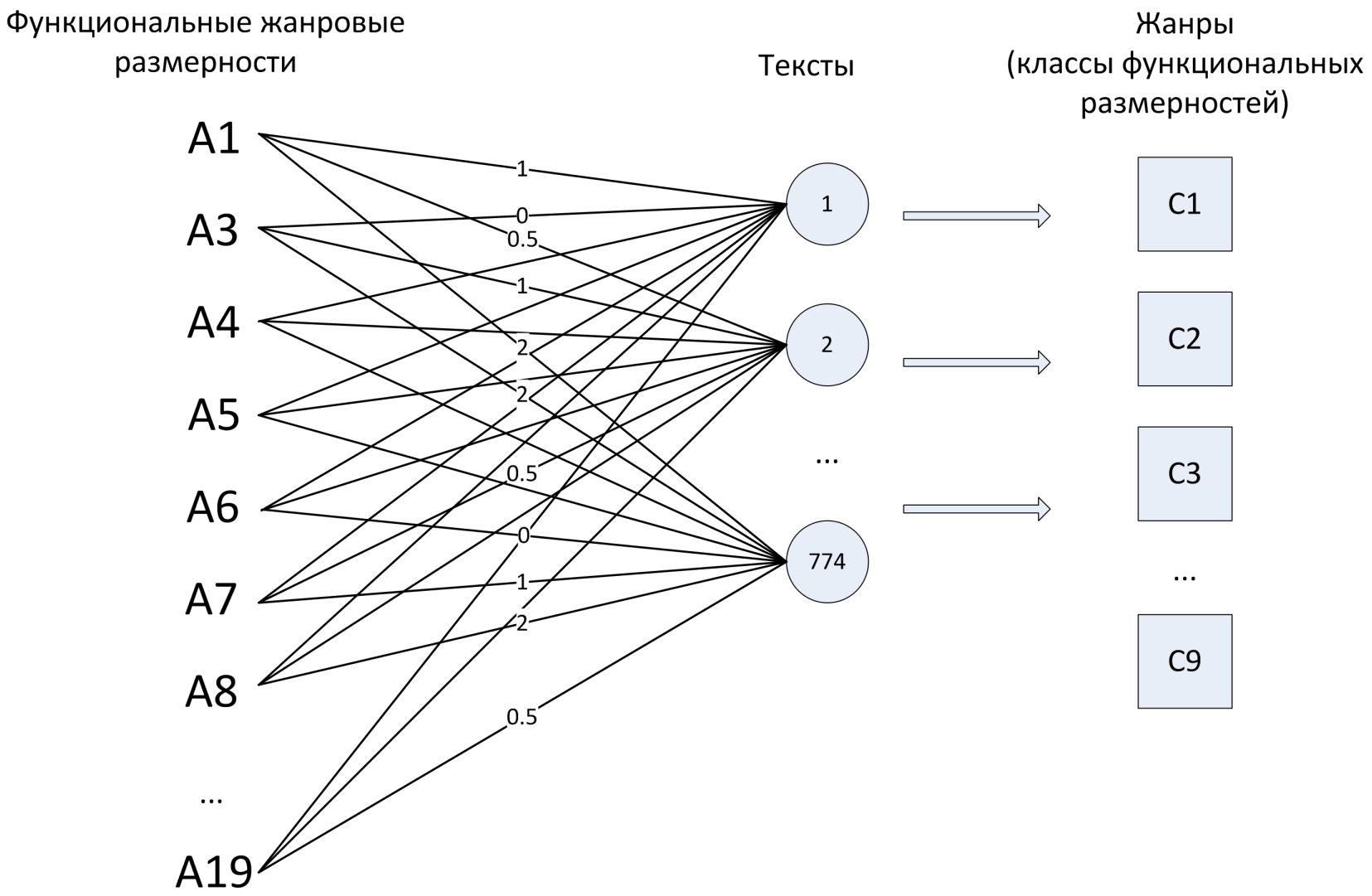
A8: hardnews До какой степени текст является информационным сообщением о недавних событиях относительно времени его написания?

A17: eval До какой степени текст содержит оценку (одобрение, критика и т.д.) того, о чем идет речь?

Оценки разметчиков:

0	ни в какой степени;	(Forsyth, Sharoff, 2014)
.5	слегка;	
1	частично;	
2	большей частью.	

Разметка текстов по Функциональным жанровым размерностям



Псевдожанры и их функциональные размерности (Lagutin et al., 2015)

Жанр	**Функциональные размерности**
Аргументативные тексты	A1, A13
Личные блоги	A11
Новостные тексты	A8
Юридические тексты	A9
Рекламные тексты	A12
Научные тексты	A14
Энциклопедические тексты	A16, -A14
Инструктивные тексты	A7
Художественные тексты	A4, A16

Отбор языковых признаков

Языковые признаки

D. Biber	MDRus Analyser
Agentless passives	Предложения с пассивными конструкциями без выраженного агента
Attributive adjectives	Адъективное определение
Average word length	Средняя длина слова
By-passives	Предложения с пассивными конструкциями с выраженным агентом
Demonstratives	Указательные местоимения *это*, *этот*, *тот*, в качестве определителей именных групп
First person pronoun	Местоимения первого лица
Indefinite pronoun	Неопределенные местоимения
Independent clause coordination	Сочинение клауз
Necessity modals	Модальные слова со значением необходимости
Nominalizations	Отглагольные существительные
Past tense	Прошедшее время
Perfect aspect	Совершенный вид
Phrasal coordination	Сочинение фразовых категорий
Total prepositional phrases	Указательные местоимения *это*, *этот*, *тот*, в качестве определителей именных групп
Place adverbs	Наречия места
Possibility modals	Модальные слова со значением возможности
Predicative adjectives	Прилагательное в предикативной функции
Private verbs	Ментальные глаголы
Public verbs	Глаголы речи
Second person pronoun	Местоимения второго лица
Split infinitives	Инфинитив
Suasive verbs	Изъяснительные придаточные
That verb complements	Все существительные
Total other nouns	Относительные придаточные клаузы
Wh-clauses	...
...	

Языковые признаки в наборе Д. Байбера, от которых мы отказались

- Amplifiers (*absolutely, completely, perfectly, strongly, thoroughly*)
- *Be* as a main verb
- Conjuncts (*consequently, e.g., furthermore, hence, i.e., nonetheless, therefore*)
- Contractions (*n't*)
- Downtoners (*almost, hardly, merely, nearly, only, partly, practically, scarcely*)
- Emphatics (*just, really, most, more, real + adjective, so + adjective*)
- Existential *there*

Языковые признаки в наборе MDRus Analyser, которые мы добавили

- Возвратные местоимения
- Указательные местоимения (*это, этот, тот*) как определители именных групп.
- Указательные местоимения (*это, этот, тот*) как проформы именных групп.
- Ментальные глаголы
- Глаголы речи
- Восклицательные предложения
- Повелительное наклонение
- Условное наклонение

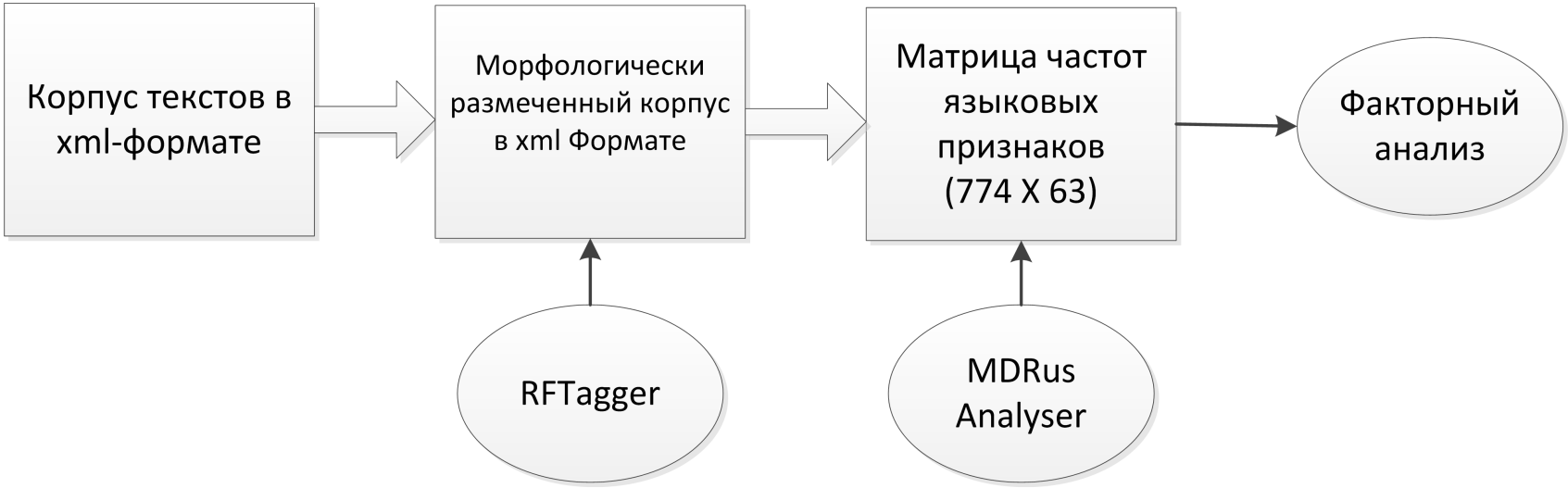
Языковые признаки в наборе MDRus Analyser, которые мы добавили

- Оценочная лексика
- Эвиденциальность (*по-видимому, говорят, по слухам, как пишут, по-моему* и др.)
- Вводные слова и конструкции отношения и оценки (*странно, увы, удивительно, надеюсь*)
- Вводные слова и конструкции отношения какого-то отрезка речи к последовательности мысли (*вдобавок, во-первых, притом*)
- Вводные слова и конструкции со значением выделения сказанного (*подчеркиваю, вообразите, заметь, например* и др.)
- Наречия степени

Языковые признаки в наборе MDRus Analyser, которые мы добавили

- Частицы (*только, лишь, хоть, же, ведь, нет-нет да и*)
- Существительные с количественным значением
- Существительные времени
- Числительные
- Глаголы существования
- Глаголы изменения признака или состояния
- Глаголы движения
- Прилагательные, характеризующие физические свойства объекта
- Прилагательные со значением времени
- Прилагательные со значением размера

Поиск языковых признаков



Инструменты

- Программа MDRus Analyser (https://github.com/Askinkaty/MDRus_analyser)
- RFTagger (Schmid and Laws, 2008) для морфологической разметки (http://www.cis.uni-muenchen.de/~schmid/tools/RFTagger/)
- НКРЯ (http://www.ruscorpora.ru)
- ГИКРЯ (GICR) (http://webcorpora.ru)
- R (для реализации факторного анализа)

Пример работы правила

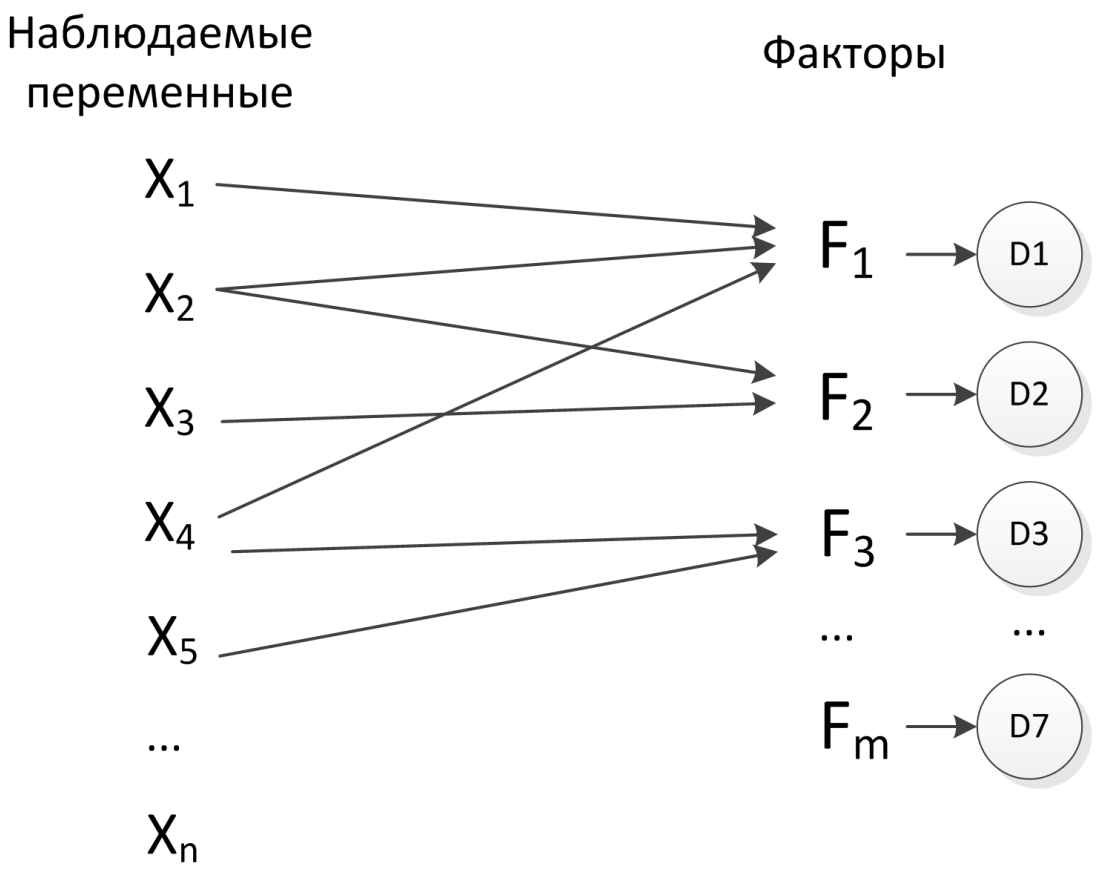
```
if d_el.get('pos') == 'V' and ('m'
== d_el.get('voice')...
```

(девушка) улыбается
(он) умылся
он возвращается
(данные) исследуются
учеными
(щеки) румянятся
морозом
друзья ссорятся

```
if (d_next_el.get('pos') == 'N' or
d_next_el.get('pos') == 'P')
and d_next_el.get(
'case') == 'i'...
```

(данные) исследуются
учеными
(щеки) румянятся
морозом

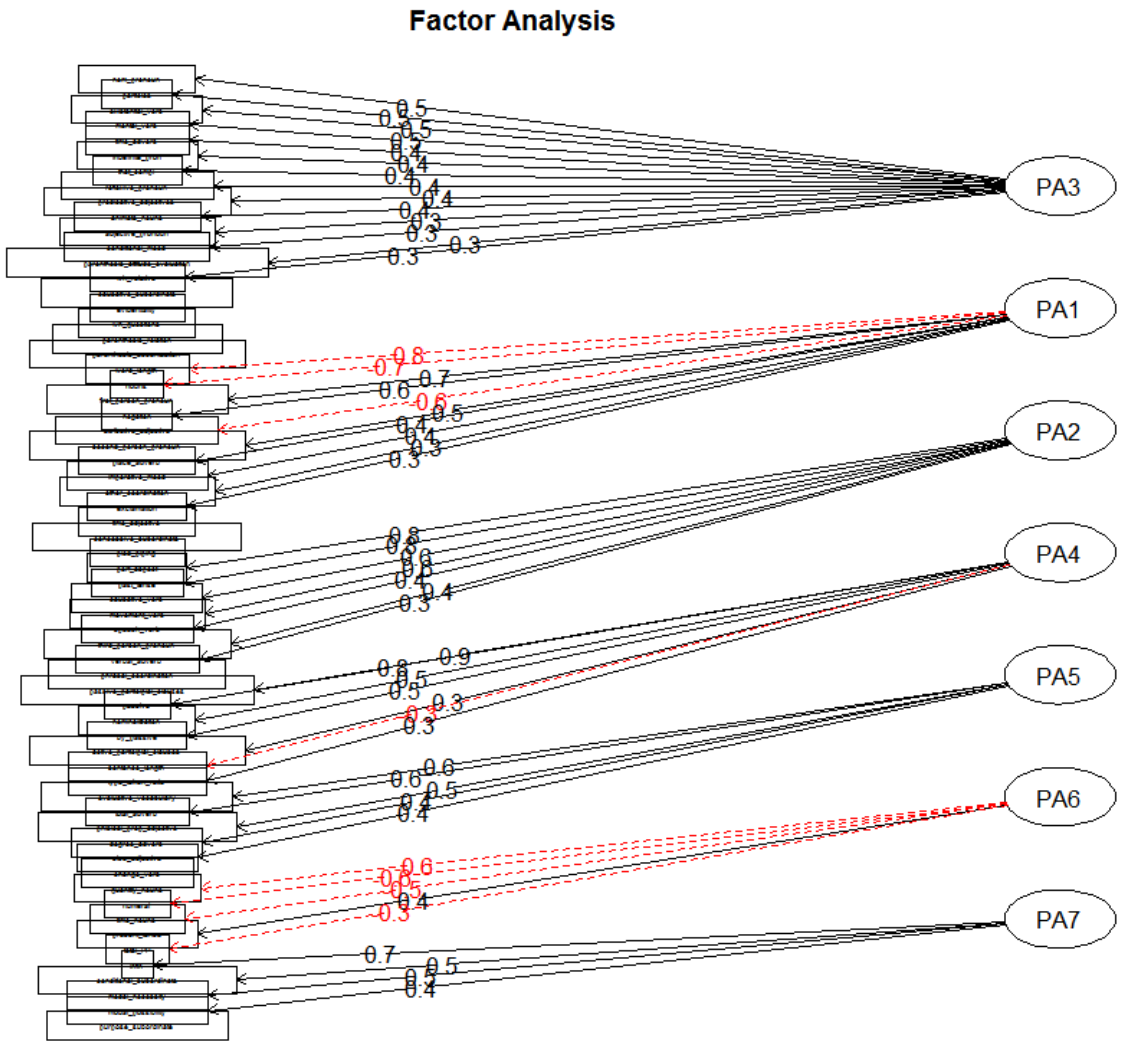
Факторный анализ



$$X_i = a_1F_1 + a_2F_2 + a_3F_3 + \ldots + a_mF_m + d_iU_i$$

$$m < n$$

Признаки и факторы



Измерения вариативности языковых признаков

- аргументативное vs. информативное (D1)
- интерактивное vs. информативное (D2)
- нарративное vs. ненарративное (D3)
- абстрактное vs. неабстрактное (D4)
- оценочное vs. неоценочное (D5)
- неколичественное vs. количественное (D6)
- директивное vs. недирективное (D7)

Примеры измерений вариативности языковых признаков

D2: интерактивное vs. информативное

Положительные признаки

- Местоимения первого лица
- Отрицание
- Местоимения второго лица
- Все наречия
- Частицы
- Ментальные глаголы
- Наречия места
- Неопределенные местоимения
- Повелительное наклонение
- Инфинитив
- Наречия времени
- Глаголы речи
- Оценочная лексика
- Сочинение клауз
- Каузативные глаголы
- Глаголы движения
- Восклицательные предложения

Примеры измерений вариативности языковых признаков

D2: интерактивное vs. информативное

(3) “— *Ты всегда врешь! Высечь тебя нужно, свиненка этакого! Я тебе уши оборву! — Да ты* что *бранишься*? — *визжит* Петя. — Что *ты* ко *мне* пристал, *дурак*? *Я* никого *не* трогаю, *не* шалю, слушаюсь, а *ты*... *сердишься*! *Ну*, за что *ты меня бранишь*? Мальчик *говорит убедительно* и так *горько* плачет, что Зайкину становится *совестно*. «И правда, за что *я* к нему придираюсь?» — *думает* он. — *Ну*, будет... будет, — *говорит* он, *трогая* мальчика за плечо. — Виноват, Петюха... прости. *Ты* у *меня умница, славный*, *я тебя* люблю”.

Примеры измерений вариативности языковых признаков

D2: интерактивное vs. информативное

Отрицательные признаки

- Средняя длина слова
- Существительные
- Адъективные определения
- Отглагольные существительные
- Пассивные конструкции без агента
- Страдательные причастные обороты
- Действительные причастные обороты
- Пассивные конструкции с выраженным агентом

Примеры измерений вариативности языковых признаков

D2: интерактивное vs. информативное

(4) "*Ввоз* в *Российскую Федерацию* и *вывоз* из *Российской Федерации отдельных товаров* и *транспортных средств* могут *быть запрещены* исходя из *соображений государственной безопасности, защиты общественного порядка, нравственности населения, жизни* и *здоровья человека, защиты животных* и *растений, охраны окружающей природной среды, защиты художественного, исторического* и *археологического достояния народов Российской Федерации* (...) *Указанные товары* и *транспортные средства подлежат соответственно немедленному вывозу* за *пределы территории Российской Федерации* либо *возврату* на *территорию Российской Федерации*, если не *предусмотрена конфискация* этих *товаров* и *транспортных средств*".

Примеры измерений вариативности языковых признаков

D5: оценочное vs. неоценочное

Положительные признаки

- Оценочная лексика
- Все наречия
- Прилагательные, характеризующие физические свойства объектов
- Наречия степени
- Прилагательные со значением размера
- Длина предложений

Примеры измерений вариативности языковых признаков

D6: неколичественное vs. количественное

Отрицательные признаки

- Существительные с количественным значением
- Числительные
- Существительные со значением времени

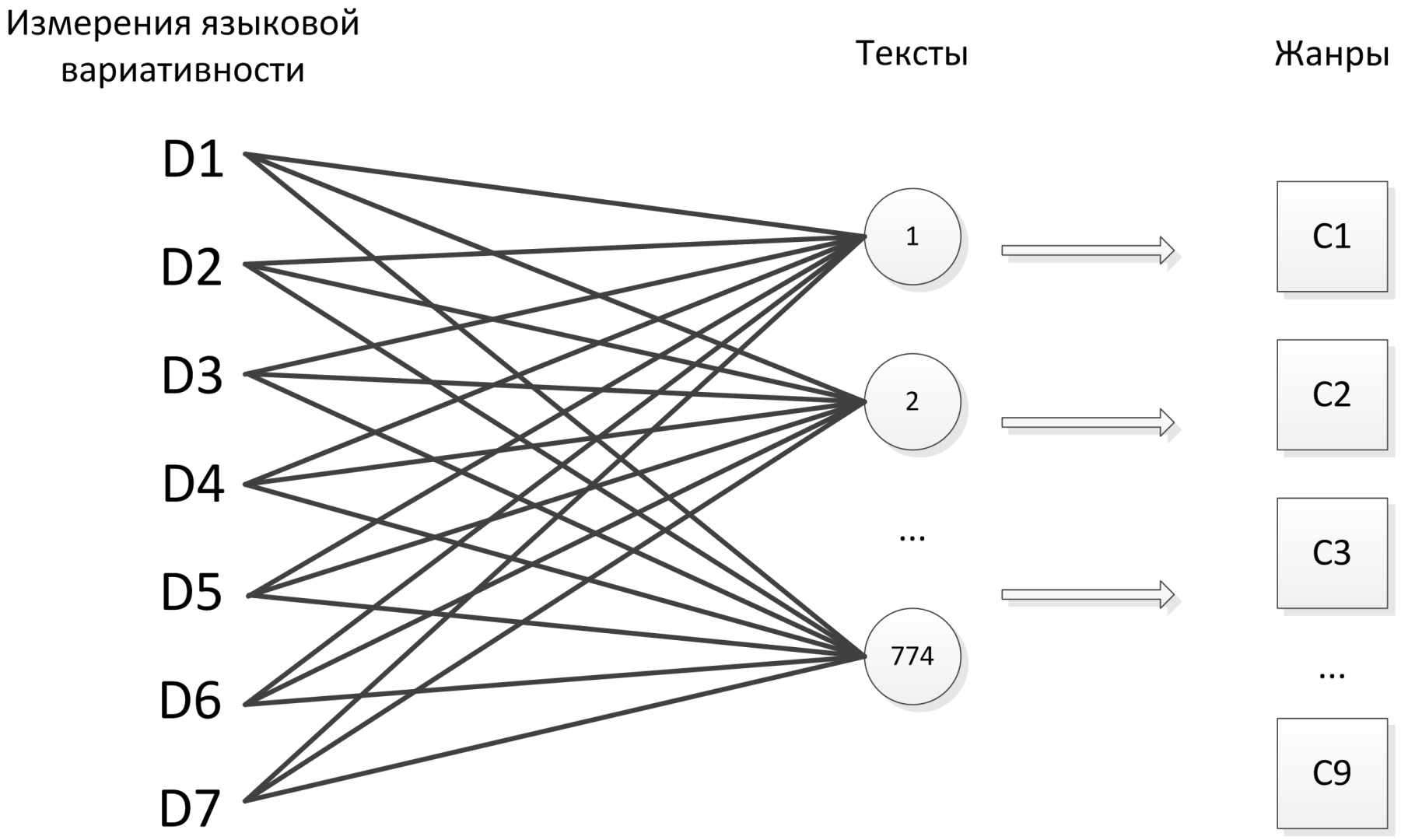
Примеры измерений вариативности языковых признаков

D7: директивное vs. недирективное

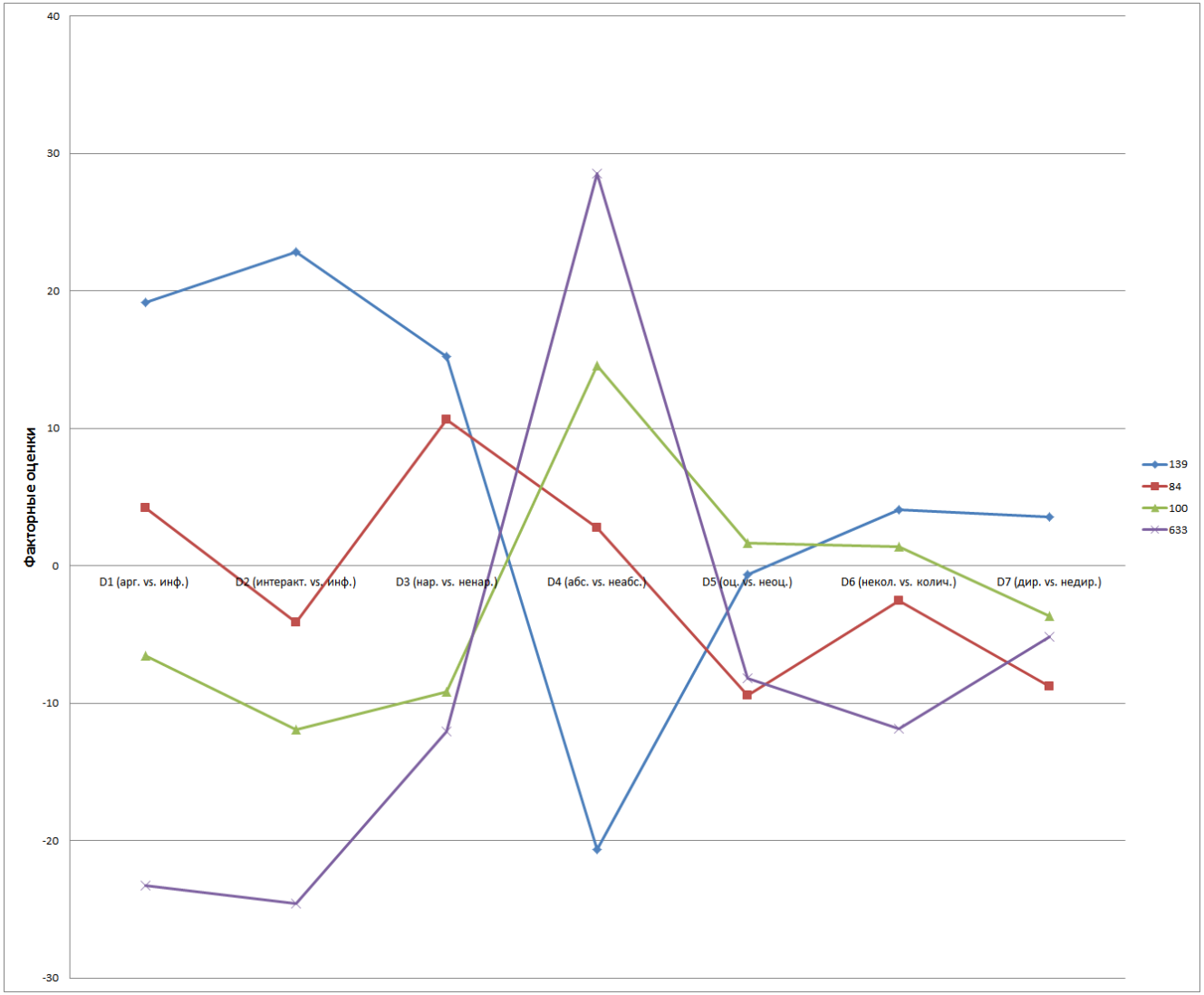
Положительные признаки

- Инфинитив
- Придаточные условия
- Модальные слова со значением необходимости
- Модальные слова со значением возможности
- Повелительное наклонение
- Местоимения второго лица
- Оценочная лексика
- Отрицание
- Оценочная лексика

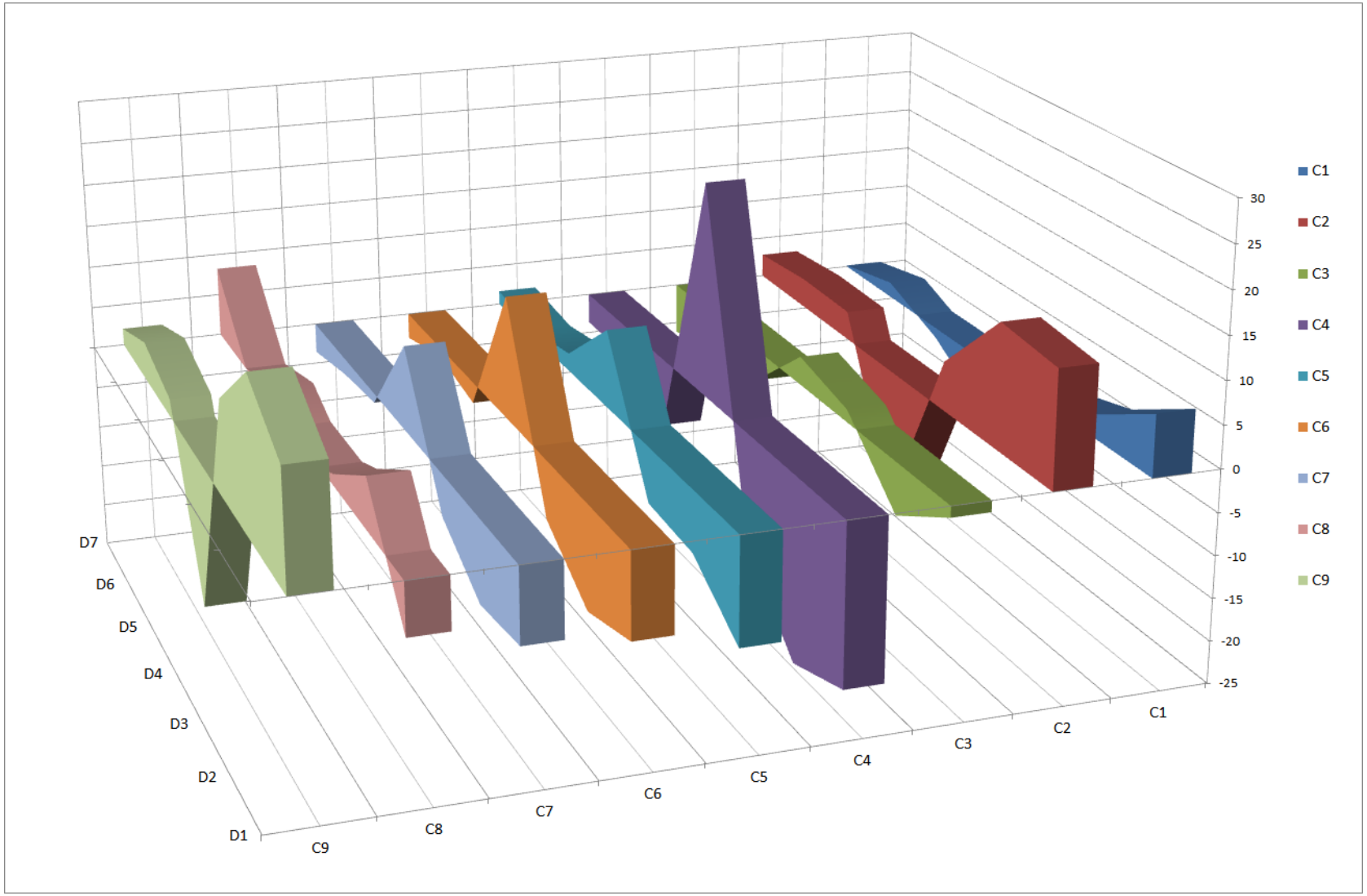
Признаки и факторы



Тексты в пространстве измерений языковой вариативности



Жанры в пространстве измерений языковой вариативности



Выводы и наблюдения

- измерение «оценочность» позволяет различать сходные по факторным профилям аргументативные тексты и блоги;
- максимальной для личных блогов является оценка по измерению «интерактивность»;
- поиск отзывов и рецензий по высоким оценкам по измерению D5 («оценочность vs. неоценочность») в личных блогах;
- новостные сообщения и новостные аналитические сообщения различаются по значениям оценок в измерении D1 («аргументативность vs. информативность»);

Выводы и наблюдения

- два поджанра рекламных текстов: неформальные объявления с сайтов знакомств и информативные, технически сложные описания рекламируемых товаров (например, автомобиля, синтезатора и т.д.);
- технические инструкции и советы в составе инструктивных текстов ;
- нейтральные по директивному измерению D7 («директивность vs. недирективность») медицинские инструкции и инструкции по охране труда;
- поиск стихов и песен в составе художественных текстов по низким оценкам в измерении D1 («аргументативность vs. информативность»).

Общие результаты анализа

- Семь измерений вариативности языковых признаков.
- Русский язык не показал значительных отличий от других языков в контексте Многомерного анализа.
- Классы функциональных жанровых размерностей и измерения языковой вариативности взаимосвязаны.
- Отклонения от среднего в факторных оценках текстов позволяют искать новые поджанры или ошибки в аннотации.
- + Функциональные размерности текстов могут быть предсказаны с точностью около 75% с помощью подмножества языковых признаков, описанных в данном исследовании (Lagutin et al., 2015).

НО! необходимо проверить полученные выводы на большем объеме текстов.

Планы на будущее

1. Повторить эксперимент по поиску и интерпретации измерений вариативности языковых признаков на большем объеме текстов разных жанров с целью проверки их универсальности, а также поиска новых измерений.
2. Исследовать вариативность признаков в отдельных фрагментах неоднородных с жанровой точки зрения текстах с целью изучения их дискурсивной структуры и поиска признаков для их автоматической классификации.

Спасибо!